T.C.
BURSA TEKNİK ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

SALDIRIDAN HABERDAR KONUŞMACI DOĞRULAMA İÇİN KARAR MALİYETİ TABANLI BİR ÖĞRENME YAKLAŞIMI

DOKTORA TEZİ

Oğuzhan KURNAZ

Elektrik-Elektronik Mühendisliği Anabilim Dalı

Elektrik-Elektronik Mühendisliği Doktora Programı

MAYIS 2025

T.C.
BURSA TEKNİK ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

SALDIRIDAN HABERDAR KONUŞMACI DOĞRULAMA İÇİN KARAR MALİYETİ TABANLI BİR ÖĞRENME YAKLAŞIMI

DOKTORA TEZİ

Oğuzhan KURNAZ
(182331541003)
ORCID: 0000-0002-4595-8031

Elektrik-Elektronik Mühendisliği Anabilim Dalı
Elektrik-Elektronik Mühendisliği Doktora Programı

Danışman: Prof. Dr. Cemal HANİLÇİ
ORCID: 0000-0002-9174-0367

MAYIS 2025

BTÜ, Lisansüstü Eğitim Enstitüsü'nün 182331541003 numaralı Doktora Öğrencisi Oğuzhan KURNAZ, ilgili yönetmeliklerin belirlediği gerekli tüm şartları yerine getirdikten sonra hazırladığı "SALDIRIDAN HABERDAR KONUŞMACI DOĞRULAMA İÇİN KARAR MALİYETİ TABANLI BİR ÖĞRENME YAKLAŞIMI" başlıklı tezini aşağıda imzaları olan jüri önünde başarı ile sunmuştur.

Tez Danışmanı :	**Prof. Dr. Cemal HANİLÇİ** Bursa Teknik Üniversitesi	.....
Jüri Üyeleri :	**Prof. Dr. Hakan GÜRKAN** Bursa Teknik Üniversitesi	.....
	Doç. Dr. Osman BÜYÜK İzmir Demokrasi Üniversitesi	.....
	Doç. Dr. Ersen YILMAZ Bursa Uludağ Üniversitesi	.....
	Dr. Öğr. Üyesi **Selma YILMAZYILDIZ KAYAARMA** Bursa Teknik Üniversitesi	.....

Teslim Tarihi :
Savunma Tarihi : 15 Mayıs 2025

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

Bu tez, 1059B142300330 numaralı TÜBİTAK 2214-A projesi ile desteklenmiştir.

İNTİHAL BEYANI

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

Öğrencinin Adı Soyadı: Oğuzhan KURNAZ

İmzası:

Eşime ve kızıma,

ÖNSÖZ

Bu tez çalışmasını hazırlarken bana destek olan ve katkı sağlayan herkese sonsuz teşekkürlerimi sunmak isterim.

University of Eastern Finland'da geçirmiş olduğum 1 yıllık süre boyunca tarafıma sağlanan 2214-A bursundan ötürü TÜBİTAK'a teşekkür ederim.

Öncelikle, bu süreçte bilgi ve deneyimleriyle beni yönlendiren, sabrı ve desteğiyle her zaman yanımda olan ve yeri geldiğinde danışmandan öte bir abilik yapan danışman hocam Prof. Dr. Cemal Hanilçi'ye sonsuz teşekkür ederim. Ayrıca, University of Eastern Finland'da gerçekleştirdiğim çalışmalar sırasında kıymetli rehberliği ve teşvik edici katkıları için Prof. Dr. Tomi Henrik Kinnunen'e ve sağladığı değerli katkılar için Dr. Jagabandhu Mishra'ya şükranlarımı sunarım.

Bu yolculukta beni yalnız bırakmayan, bilgi ve tecrübeleriyle destek olan Bursa Teknik Üniversitesi'ndeki değerli hocalarıma ve meslektaşlarıma, ayrıca üniversitedeki tüm arkadaşlarıma teşekkür ederim.

Bugünlere gelmemde en büyük pay sahibi olan sevgili annem, babam ve kardeşime, bana her zaman inandıkları ve destek oldukları için minnettarım.

Son olarak, bu zorlu süreçte yanımda olan, desteğini hiçbir zaman esirgemeyen sevgili eşime ve varlığıyla bana huzur veren biricik kızıma en içten teşekkürlerimi sunarım. Bu tez, onların sevgi ve desteği olmadan mümkün olamazdı.

Mayıs 2025 Oğuzhan KURNAZ

İÇİNDEKİLER

Sayfa

KISALTMALAR

a-DCF : Architecture-agnostic Detection Cost Function
ASV : Automatic Speaker Verification
CNN : Convolutional Neural Network
CM : Countermeasure
DF : Deepfake
DNN : Deep Neural Network
EER : Equal Error Rate
FAR : False Acceptance Rate
FRR : False Rejection Rate
GAN : Generative Adversarial Network
GMM : Gaussian Mixture Model
HMM : Hidden Markov Model
JFA : Joint Factor Analysis
LA : Logical Access
LR : Logistic Regression
MAP : Maximum a Posteriori
MFM : Max Feature Map
PA : Physical Access
ROC : Receiver Operating Characteristic
SASV : Spoofing-Aware Speaker Verification
SVM : Support Vector Machine
TDNN : Time Delay Neural Network
TTS : Text-to-Speech
UBM : Universal Background Model
VC : Voice Conversion
VQ : Vector Quantization

SEMBOLLER

$\boldsymbol{\lambda_g}$	**:** Gerçek sese ait model parametreleri
$\boldsymbol{\lambda_s}$	**:** Sahte sese ait model parametreleri
$\boldsymbol{\theta}$	**:** Eşik değeri
$\boldsymbol{\phi}(.)$	**:** Karar skoru
$\boldsymbol{s_{ASV}}$	**:** ASV Skoru
$\boldsymbol{s_{CM}}$	**:** CM Skoru
$\boldsymbol{e_{ASV}^{tst}}$	**:** ASV Test Embedding
$\boldsymbol{e_{ASV}^{enr}}$	**:** ASV Enrollment Embedding
$\boldsymbol{e_{CM}^{tst}}$	**:** CM Test Embedding
$\boldsymbol{C}$	**:** Maliyet
$\boldsymbol{P}$	**:** Hata Oranı
$\boldsymbol{\pi}$	**:** Konuşmacı Olasılıkları
$\boldsymbol{\tau}$	**:** Karar Eşiği
$\boldsymbol{L}$	**:** Kayıp Fonksiyonu

ÇİZELGE LİSTESİ

Sayfa

ŞEKİL LİSTESİ

Sayfa

SALDIRIDAN HABERDAR KONUŞMACI DOĞRULAMA İÇİN KARAR MALİYETİ TABANLI BİR ÖĞRENME YAKLAŞIMI

ÖZET

Otomatik konuşmacı doğrulama (Automatic Speaker Verification – ASV) sistemleri, biyometrik kimlik doğrulama alanında hem kullanıcı dostu hem de etkili bir doğrulama aracı olarak önemli bir rol oynamaktadır. Bu sistemler, konuşmacı kimliğini belirlemek veya doğrulamak amacıyla özellikle güvenlik uygulamalarında, erişim kontrol sistemlerinde ve mobil cihaz doğrulamalarında yaygın şekilde kullanılmaktadır. Ancak bu sistemler, yanıltma saldırılarına karşı savunmasız olduğundan güvenlik riskleri barındırmaktadır. Özellikle, tekrar oynatma saldırıları (replay attacks), metinden konuşma sentezleme (text-to-speech synthesis) ve ses dönüştürme (voice conversion) gibi yanıltma saldırıları, konuşmacı doğrulama sistemlerinin güvenliğini tehlikeye atmaktadır. Bu risklere karşı koymak için, saldırı tespit ve önleme yeteneğine sahip çeşitli yanıltma saldırısı tespit (countermeasure - CM) sistemler geliştirilmiştir.

Son yıllarda, ASV ve CM sistemlerinin entegrasyonu ile yanıltma saldırılarından haberdar konuşmacı doğrulama sistemlerinin geliştirilmesi çalışmaları yaygınlaşmıştır. Bu yeni nesil sistemler, konuşmacı doğrulama süreçlerini yalnızca kullanıcı doğrulama değil, aynı zamanda yanıltma saldırılarına karşı dayanıklılık açısından da güçlendirmektedir. Bu tez, saldırıdan haberdar konuşmacı doğrulama sistemlerinin geliştirilmesine odaklanmaktadır. Özellikle, çok aşamalı birleştirme yöntemleri, derin öznitelik seviyesinde işlem yapan paralel mimariler ve metrik tabanlı optimizasyon yöntemleri kullanılmıştır.

Çalışmada, ECAPA-TDNN, WavLM ve AASIST gibi öncü modeller entegre edilerek hem skor seviyesi hem de derin öznitelik seviyesinde birleştirme teknikleri uygulanmıştır. Bunun yanı sıra, mimariden bağımsız bir değerlendirme metriği olan a-DCF, kayıp fonksiyonu olarak optimize edilmiştir. Bu kapsamda, ASVspoof 2019 ve ASVspoof 5 veri kümeleri üzerinde yapılan kapsamlı deneyler, önerilen yöntemlerin etkinliğini göstermiştir. Elde edilen sonuçlar, hata oranlarında belirgin düşüşler sağlarken, hedef örnekler ile hedef olmayan ve sahte örnekler arasında daha iyi bir ayrım yapılmasını mümkün kılmıştır.

Bu tez, mevcut en ileri modeller ve optimizasyon tekniklerinin bir araya getirilmesiyle, biyometrik güvenliğin daha ileri taşınmasına katkıda bulunmaktadır. Özellikle, gerçek dünya uygulamalarında güvenliğin artırılması açısından önemli bir adım atılmıştır. Bu bağlamda, çalışma hem akademik literatüre hem de pratik uygulamalara yönelik değerli bir katkı sunmaktadır.

Anahtar kelimeler: Konuşmacı doğrulama, Yanıltma saldırısı tespiti, Saldırıdan haberdar konuşmacı doğrulama

A DECISION COST-BASED LEARNING APPROACH FOR SPOOFING-AWARE SPEAKER VERIFICATON

SUMMARY

Automatic Speaker Verification (ASV) systems play a significant role in biometric identity verification, offering both user-friendly and effective authentication solutions. These systems are widely utilized in various applications, including security systems, access control, and mobile device authentication, with the primary goal of identifying or verifying speaker identity. However, ASV systems remain vulnerable to spoofing attacks, thereby posing significant security risks in the absence of proper countermeasures. Specifically, attacks such as replay attacks, text-to-speech synthesis, and voice conversion threaten the reliability of speaker verification systems. To mitigate these threats, various countermeasure (CM) systems have been developed to detect and prevent spoofing attempts.

In recent years, there has been a growing body of research on developing spoofing-aware speaker verification systems through the integration of ASV and CM systems. These next-generation systems enhance speaker verification by not only verifying user identity but also improving resilience against spoofing attacks. In particular, multi-stage fusion strategies, parallel architectures operating on deep feature representations, and metric-based optimization methods have been employed.

In this study, state-of-the-art models such as ECAPA-TDNN, WavLM, and AASIST are integrated, and both score-level and deep embedding-level fusion techniques are employed. Additionally, a system-independent evaluation metric, the architecture-agnostic detection cost function (a-DCF), is optimized as the loss function. Comprehensive experiments conducted on the ASVspoof 2019 and ASVspoof 5 datasets demonstrate the effectiveness of the proposed methods. The results show significant reductions in error rates and enable more accurate differentiation between target, nontarget and spoof speech samples.

This thesis contributes to advancing biometric security by combining cutting-edge models and optimization strategies. It represents a significant step toward enhancing the security of real-world applications. Accordingly, this work offers valuable contributions to both academic research and practical implementations in the field of speaker verification.

Keywords: Automatic speaker verification, Spoofing countermeasure, Spoofing-robust speaker verification

1. GİRİŞ

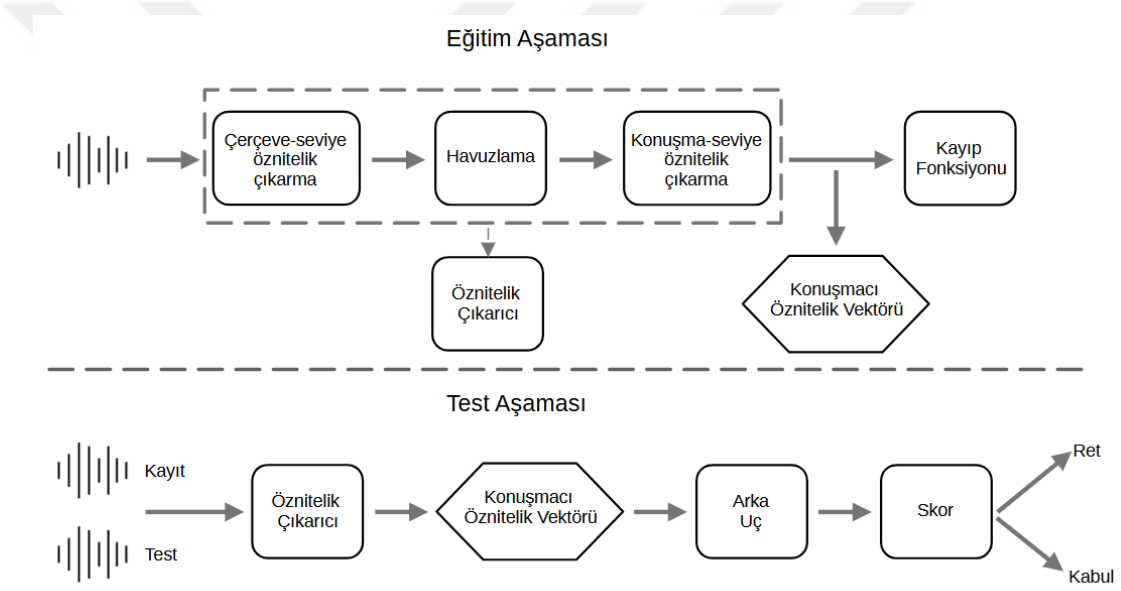
Bu tez, otomatik konuşmacı doğrulama (Automatic Speaker Verification - ASV) sistemleriyle entegre edilerek daha güvenilir ve gürbüz bir kimlik doğrulama sistemi oluşturmayı hedefleyen, etkili ve anlaşılabilir yanıltma saldırısı karşıtı önlemlerin geliştirilmesine odaklanmaktadır. Konuşmacı tanıma alanı, yıllar içinde yüksek doğrulukta modeller geliştirme konusunda önemli ilerlemeler kaydetmiş olsa da ASV sistemlerinin yanıltma saldırılarına karşı savunmasızlığı, son yıllarda giderek daha fazla dikkat çekmektedir. Bu ilginin artmasında, 2010'lu yıllardan itibaren derin öğrenme alanındaki gelişmelerin büyük etkisi olmuştur. Bu gelişmeler, yalnızca yapay olarak üretilen konuşmaların incelenmesini hızlandırmakla kalmamış, aynı zamanda bu tür konuşmaların ASV sistemlerine oluşturduğu potansiyel riskleri de gözler önüne sermiştir. Bu zorluklara yanıt olarak, yalnızca bu tehditleri tespit edip etkisiz hale getiren değil, aynı zamanda yanıltma saldırısı karşıtı önlemlerin çalışma mekanizmalarına da açıklık getiren sistemlerin geliştirilmesi büyük önem taşımaktadır.

1.1 Otomatik Konuşmacı Doğrulama

ASV sistemleri, bir kişiye ait konuşma sinyalini kullanarak kimlik doğrulamasını gerçekleştiren biyometrik sistemlerdir [1]. Bu sistemler, konuşmanın fizyolojik ve davranışsal özelliklerini - örneğin ses üretim mekanizması, burun boşluğu, akciğer hava basıncı ve tonlama [2] gibi - analiz ederek konuşma modellemesi gerçekleştirerek konuşmaları birbirinden ayırt etmeyi sağlamaktadır.

Bir ASV sistemi *eğitim* ve *test* olmak üzere iki aşamadan oluşur. Şekil 1.1'de bu aşamalar görselleştirilmiştir. Şeklin üst kısmında yer alan eğitim aşaması, bilinen bir konuşmacı grubundan alınan eğitim konuşmalarına bir konuşmacı etiketinin atandığı çok sınıflı bir sınıflandırma işlemi gerçekleştirmek üzere bir derin sinir ağı modelinin eğitildiği aşamadır. Bu aşamada, Zaman Gecikmeli Sinir Ağları (Time Delay Neural Network - TDNN) [3, 4] veya evrişimsel sinir ağları (Convolutional Neural Network – CNN) gibi sınıflandırıcılar, konuşma sinyalinden konuşmacıya özgü öznitelikleri

çıkarmak üzere eğitilir. Bu ağlar, konuşma sinyalinin kısa çerçevelerindeki spektral zarf içindeki akustik örüntüleri yakalamak üzere eğitilmektedir. Daha sonra çıkarılan bu çerçeve düzeyindeki öznitelikler, konuşmanın zamansal dinamiklerini özetleyen, konuşma düzeyinde öznitelikler elde etmek için bir havuzlama katmanından geçirilir [4, 5]. En son aşamada elde edilen bu öznitelikler derin öznitelik vektörleri (deep speaker embedding) olarak adlandırılmakta olup, konuşmacının kimliğini kompakt bir biçimde temsil eden ve test aşamasında kullanılan temel yapı taşlarıdır. Derin öznitelik vektörlerini elde etmek için eğitilen ağ parametrelerini optimize etmek için bir kayıp fonksiyonu kullanılır; bu fonksiyon, aynı konuşmacıya ait öznitelikler arasındaki mesafeyi en aza indirirken, farklı konuşmacılara ait öznitelikler arasındaki mesafeyi en üst düzeye çıkaracak şekilde tasarlanmaktadır [6, 7].



Şekil 1.1. Genel bir konuşmacı doğrulama sistemi [8].

Şekil 1.1'in alt kısmında gösterilen test aşamasında ise, test konuşma sinyali (test utterance) ile iddia edilen konuşmacıya ait eğitim konuşma sinyali (enrollment utterance) öznitelikleri karşılaştırılır. Her iki konuşma sinyali, eğitim aşamasında eğitilen aynı öznitelik çıkarıcıdan geçirilerek derin konuşma öznitelik vektörleri elde edilir. Daha sonra bir arka uç (back-end) sistemi aracılığı ile eğitim ve test sinyallerinden elde edilen öznitelik vektörlerinin benzerliği değerlendirilerek bir karar skoru hesaplanır. Hesaplanan karar skoru bir eşik değer ile kıyaslanarak kimlik iddiası kabul veya reddedilir. Yüksek bir skor, her iki konuşma sinyalinin aynı konuşmacıya ait olduğunu belirterek kimlik iddiasının kabul edilmesini sağlarken düşük bir skor ise,

konuşmaların farklı konuşmacılara ait olduğunu göstererek kimlik iddiasının reddedilmesine olanak sağlamaktadır [8].

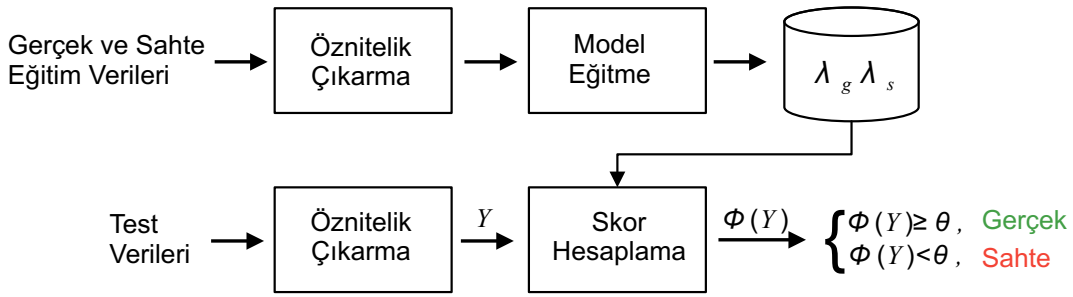
1.2 Yanıltma Saldırısı Tespiti

Literatürdeki birçok çalışmada, ASV sistemlerinin farklı şekillerde gerçekleştirilen yanıltma saldırılarına karşı savunmasız olduğu gösterilmiştir [10]. ASV sistemlerinin yanıltma saldırılarına karşı savunmasızlığı önemli güvenlik endişelerine yol açmaktadır. Yanıltma saldırıları genellikle hedef konuşmacının ses özelliklerini taklit etmeyi amaçlamaktadır. Bu saldırılar, basitçe daha önce kaydedilmiş bir konuşma kaydının tekrar oynatılması [11] ya da daha teknik yöntemlerle, örneğin metinden konuşma sentezi (TTS) veya ses dönüştürme (VC) teknolojileriyle [12, 13] gerçekleştirilebilmektedir. TTS ve VC teknolojileri, hedef konuşmacının sesini taklit eden yapay konuşmalar üretmek üzere geliştirilen yöntemlerdir. Bu teknolojiler o kadar ileri seviyeye ulaşmıştır ki, artık bu sistemler ile üretilen yapay konuşmalar gerçek insan seslerinden ayırt edilemeyecek kadar gerçekçi olmaktadır.

ASV sistemlerinin eğitimi sırasında kullanılan veri tabanları genellikle yalnızca gerçek/sahte olmayan (bonafide) konuşma kayıtlarını içerdiği için, bu sistemler özenle hazırlanmış yanıltma saldırılarına karşı savunmasızdır. Bu durum ASV sistemlerinde yanlış alarm oranının artışına neden olabilmektedir [10]. Dolayısıyla, ASV sistemlerinin dayanıklılığını ve güvenilirliğini artırmak için etkili yanıltma saldırısı karşıtı önlemler (CM) geliştirilmesine ihtiyaç duyulmaktadır.

Şekil 1.2'de gösterildiği gibi, yanıltma saldırı tespiti sistemi (CM), bir konuşma kaydının gerçek mi yoksa sahte mi olduğunu belirlemek için önce gerçek ve sahte eğitim verileriyle eğitilir. Bu süreçte, sistem, gerçek ve sahte seslere ait model parametrelerini (λ_g ve λ_s) öğrenerek, bu iki sınıfı ayırt edebilecek bir model oluşturur. Test aşamasında ise, test verilerinden çıkarılan öznitelikler (Y) ve eğitim sırasında elde edilen gerçek ve sahte ses sınıflarına ait modeller birlikte kullanılarak bir karar skoru ($\Phi(Y)$) hesaplanır. Hesaplanan bu skor, belirli bir eşik değeri (θ) ile karşılaştırılarak

konuşmanın "Gerçek" ($\Phi(Y) \geq \theta$) ya da "Sahte" ($\Phi(Y) < \theta$) olarak sınıflandırılması sağlanır.



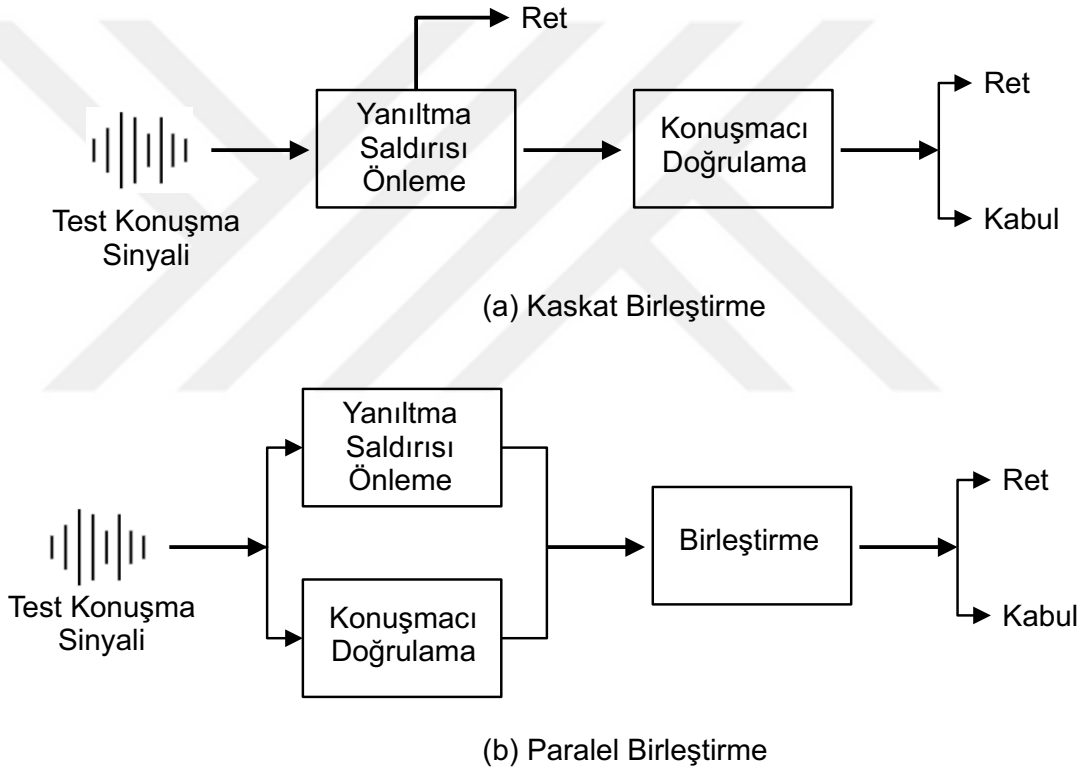
Şekil 1.2: Yanıltma saldırılarının tespiti sisteminin genel yapısı.

1.3 Saldırıdan Haberdar Konuşmacı Doğrulama

ASV sistemlerinin yanıltma saldırısına karşı savunmasızlığı, güvenilir kimlik doğrulama süreçleri için CM sistemleri ile birlikte çalışmayı zorunlu kılmaktadır. Tek başına çalışan bir ASV sistemi, genellikle yalnızca konuşmacının kimliğini doğrulama problemini çözmek üzere geliştirildiği için yanıltma saldırısının tespiti konusunda yetersiz kalmaktadır. Yanıltma saldırılarının giderek karmaşıklaşması, ASV ve CM sistemlerinin birlikte çalışmasını zorunlu hale getirmiştir. Bu iki sistemin birleştirilmesi, sadece yanıltma saldırısı tespitinde değil, aynı zamanda kimlik doğrulama süreçlerinin güvenilirliğinin artırılmasında da kritik bir rol oynamaktadır. CM sistemleri, ASV sistemine ulaşmadan önce sahte sinyalleri tespit ederken, ASV sistemleri bu verileri analiz ederek konuşmacı kimliğini doğrular. Bu entegrasyon, yanıltma saldırılarının etkisini minimize etmek ve yanlış alarm oranını düşürmek için vazgeçilmezdir.

ASV ve CM sistemlerinin birlikte çalışması, farklı entegrasyon stratejileriyle daha güvenilir doğrulama sistemleri oluşturulmasını sağlamaktadır. Şekil 1.3'te gösterildiği gibi, bu entegrasyon stratejilerinin en yaygın örnekleri arasında kaskat (cascade) ve paralel birleştirme yapıları bulunmaktadır. Kaskat birleştirme yapısında (Şekil 1.3 (a)) önce, CM sisteminin yanıltma saldırılarını tespit etmesi sağlanarak yalnızca gerçek (bonafide) olarak sınıflandırılan konuşmaların ASV sistemi tarafından işlenmesi amaçlanmaktadır. Bu modelde, CM sistemi sahte konuşmaları filtreleyerek ASV sistemine yalnızca güvenilir konuşma sinyallerinin iletilmesi sağlanmaktadır. Bu yaklaşım, sahte konuşmaların sistemden tamamen elenmesini hedefler ve bu sayede

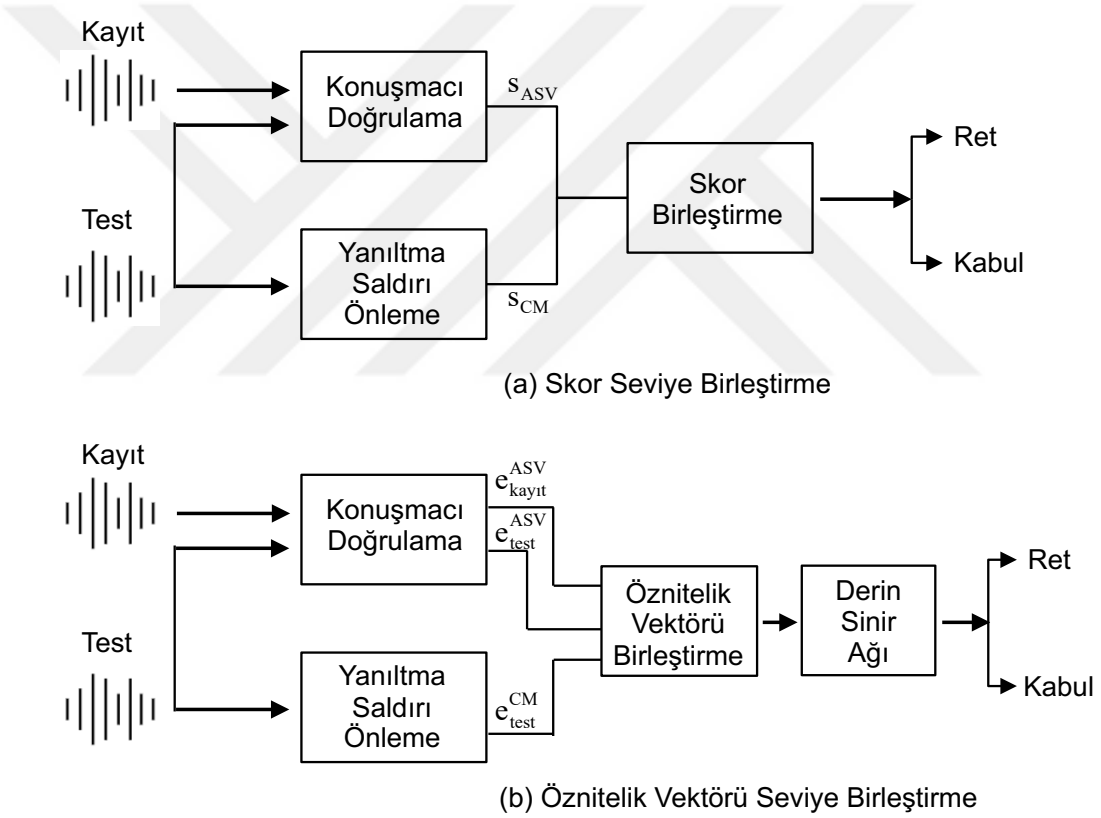
ASV sisteminin, yalnızca gerçek konuşmalar üzerinden kimlik doğrulama yapmasını sağlamaktadır. Benzer şekilde, önce ASV sisteminin ardından CM sisteminin yerleştirildiği bir kaskat yapı da tasarlanabilir. Paralel birleştirme yapısında (Şekil 1.3 (b)), CM ve ASV sistemleri eş zamanlı olarak çalışmaktadır. Bu birleştirme yönteminde, her iki sistemin çıktıları (skor veya öznitelik) birleştirilir ve nihai bir karar mekanizmasıyla "Ret" veya "Kabul" kararı verilir. Paralel birleştirme, CM ve ASV sistemlerinin bağımsız olarak yanıltma saldırısı tespiti ve konuşmacı doğrulama süreçlerine katkıda bulunmasını sağlar. Özellikle bu birleştirme yönteminde, her iki sistemin ürettiği bilgilerin birleştirilmesiyle daha güçlü bir karar mekanizması oluşturulabilmektedir.



Şekil 1.3: ASV ve CM sistemlerin birleştirilmesi [14].

Paralel yapıların daha da geliştirilmiş versiyonları, entegrasyonun farklı seviyelerde gerçekleştirilmesine olanak tanıyarak ASV ve CM sistemlerinin daha etkin bir şekilde birleştirilmesi sağlanmaktadır. Şekil 1.4'te görüldüğü gibi, bu entegrasyon yöntemleri arasında skor seviyesinde birleştirme ve öznitelik vektörü seviyesinde birleştirme yöntemleri bulunmaktadır. Şekil 1.4 (a)'da gösterilen skor seviyesinde birleştirme, CM ve ASV sistemlerinin ürettiği karar skorlarının (s_{CM} ve s_{ASV}) birleştirilmesiyle gerçekleştirilir. Skor birleştirme mekanizmasında, her iki sistem tarafından hesaplanan

karar skorları birleştirilerek "Ret" veya "Kabul" kararı verilmektedir. Bu yöntem, her iki sistemin bağımsız olarak çalışmasını sağlayarak sahte veya gerçek konuşma hipotezlerinden birine daha fazla destek sunan nihai bir karar skoru üretmektedir. Skor seviyesinde birleştirme, sistemin karar mekanizmasının basit ancak etkili bir şekilde entegre edilmesine olanak sağlamaktadır. Şekil 1.4 (b)'de gösterilen öznitelik vektörü seviyesinde birleştirmede ise, daha karmaşık bir entegrasyon yapısı bulunmaktadır. Bu yöntemde, CM ve ASV sistemlerinden elde edilen öznitelik vektörleri ($e_{\text{ASV}}^{\text{kayıt}}$, $e_{\text{ASV}}^{\text{test}}$, $e_{\text{CM}}^{\text{test}}$) birleştirilerek tek bir ortak temsil vektörü oluşturulur. Bu birleştirilmiş öznitelik vektörü, bir derin sinir ağı tarafından işlenerek "Ret" veya "Kabul" kararının verilmesi amaçlanmaktadır.



Şekil 1.4: (a) Skor seviyesinde birleştirme ve (b) öznitelik vektörü seviyesinde birleştirme [14].

Bu entegrasyon stratejileri, ASV ve CM sistemlerinin zayıflıklarını telafi ederek daha dayanıklı ve güvenilir bir konuşmacı doğrulama mekanizması oluşturur. Özellikle paralel yapıdaki entegrasyon seviyelerinin esnekliği, sistemlerin saldırılara karşı daha etkili ve uyumlu bir şekilde çalışmasını sağlar. Böylelikle, yanıltma saldırılarının tespit

edilmesi ve konuşmacı kimliğinin doğrulanması süreçlerinde bütüncül bir yaklaşım benimsenmiş olmaktadır.

1.4 Motivasyon ve Amaç

ASV ve CM sistemlerinin entegrasyonu, yanıltma saldırılarına karşı daha güvenilir ve dayanıklı konuşmacı doğrulama sistemleri oluşturmak için kritik bir gereklilik haline gelmiştir. Özellikle derin öznitelik vektörü seviyesinde ve skor seviyesinde birleştirme yöntemleri, bu entegrasyonun etkili bir şekilde gerçekleştirilmesi için öne çıkan stratejiler arasında yer almaktadır. Bu tezde, saldırıdan haberdar bir konuşmacı doğrulama sistemi (SASV) tasarımı için derin öznitelik seviyesinde ve skor seviyesinde birleştirme yöntemleri üzerine detaylı çalışmalar yapılmıştır. Ayrıca, literatürde SASV sistemlerini değerlendirme metriği olarak yaygın bir şekilde kullanılan a-DCF metriği, bu tez kapsamında bir sinir ağının optimize edilmesi için kayıp fonksiyonu olarak önerilmiştir. Bu yenilikçi yaklaşım, sistem performansını artırmaya yönelik metrik tabanlı bir optimizasyon sağlamayı amaçlamaktadır.

Bu tez çalışmasının amacı, SASV sistemlerinin yanıltma saldırılarına karşı daha etkin bir şekilde çalışmasını sağlayacak yeni yöntemler geliştirmek ve sistem performansını metrik tabanlı bir optimizasyon ile artırmaktır. Bu kapsamda aşağıdaki bilimsel katkılar sunulmuştur:

- *Skor Seviyesinde Çok Aşamalı Birleştirme Yapısı:* SASV sisteminde, skor seviyesinde birleştirme işlemleri için çok aşamalı (multi-stage) bir yapı tasarlanmış ve bu yapının yanıltma saldırı tespitindeki etkinliği detaylı bir şekilde analiz edilmiştir. Bu yaklaşım, farklı aşamalarda karar süreçlerinin optimize edilmesini sağlayarak, daha doğru ve güvenilir sonuçlar elde edilmesini hedeflemektedir. *Bu doğrultuda yapılan çalışmalar 33. Sinyal İşleme ve İletişim Uygulamaları (SIU 2025) Kurultayı'nda sunulmak üzere kabul edilmiştir.*

- *Derin Öznitelik Vektörü Seviyesinde Paralel Birleştirme:* Paralel derin öznitelik vektörü birleştirme yönteminin SASV sistemlerine etkisi araştırılmış ve bu seviyede birleştirmenin, konuşmacı ve yanıltma bilgilerinin zenginleştirilmiş bir temsille sunulmasına nasıl katkı sağladığı incelenmiştir. *2024 yılında düzenlenen ASVspoof 5 yarışması için önerilen bu yöntem, ASVspoof 2024 Workshop etkinliğinde yayınlanmıştır.*

- *a-DCF Tabanlı Kayıp Fonksiyonu:* SASV sistemlerini değerlendirmede kullanılan a-DCF metriği, bu tezde bir sinir ağının optimize edilmesi için kayıp fonksiyonu olarak önerilmiştir. Bu metrik tabanlı kayıp fonksiyonu, sistemin performansını değerlendirme metrikleriyle uyumlu bir şekilde optimize etmesine olanak tanıyarak, hedeflenen performans ölçütlerine doğrudan odaklanmıştır. *Bu kapsamda yapılan çalışmalar, IEEE Signal Processing Letters dergisinde yayınlamıştır.*

Bu motivasyon ve amaç doğrultusunda, bu tez, SASV sistemlerinin daha güvenilir, performans odaklı ve yanıltma saldırılarına karşı dirençli bir şekilde tasarlanması için yenilikçi yöntemler sunmaktadır. Bu tez çalışmasının sonuçları hem literatüre katkı sağlamayı hem de pratikte daha güvenilir konuşmacı doğrulama sistemleri geliştirilmesine rehberlik etmeyi hedeflemektedir.

2. LİTERATÜR TARAMASI

Bu bölümde, tez genelinde olduğu gibi sırasıyla ASV sistemleri, CM sistemleri ve son olarak Saldırıdan Haberdar Konuşmacı Doğrulama (SASV) sistemleri hakkında detaylı bir literatür taraması sunulacaktır.

2.1 Konuşmacı Doğrulama Üzerine Gerçekleştirilen Çalışmalar

Konuşmacı doğrulama alanı, basit spektral analiz yöntemlerinden modern sinir ağı tabanlı modellere uzanan bir evrim süreciyle, sistemlerin doğruluğunu, dayanıklılığını ve güvenilirliğini artırmayı başarmıştır. İlk çalışmalar, özellikle [15] çalışmasında gerçekleştirilen spektrogram tabanlı eşleştirme yöntemleriyle başlamış, bu çalışmalar konuşmacılar arası değişimlerin tanımlanmasında önemli rol oynamıştır. [16] çalışmasında ise konuşmacı doğrulama ve tanıma kavramlarını birbirinden ayırarak, formant frekansları ve perde (pitch) gibi özelliklerin bu sistemlerde nasıl kullanılabileceğini incelenmiştir. Bu dönem, [17] tarafından önerilen metinden bağımsız konuşmacı doğrulama yöntemleriyle genişlemiş, doğrusal öngörü (linear prediction) katsayılarıyla sistemlerin daha esnek hale gelmesi sağlanmıştır. [18] çalışmasında cepstrum tabanlı özellikler ve dinamik zaman bükme (DTW) algoritmalarının tanıtılması ise düşük kaliteli kanallarda bile doğrulama sistemlerinin dayanıklılığını artırmıştır.

Konuşma sinyallerinden elde edilen akustik özniteliklerin olasılık dağılımlarını modellemek için Vektör Nicemleme (Vector Quantization - VQ) ve Gauss Karışım Modelleri (GMM) sıklıkla kullanılmıştır [19]. VQ yaklaşımında, her konuşmacıya ait kod kitapları oluşturularak, test öznitelikleri bu kitaplarla karşılaştırılarak kimlik tespiti gerçekleştirilmektedir. Buna karşın, GMM modelleri daha esnek bir yapıya sahip olup, kovaryans matrisleri aracılığıyla belirsizlikleri daha etkin biçimde yansıtmaktadır [1].

2000'li yıllarda, konuşmacı doğrulama sistemlerinde yaygınlık kazanan Gauss Karışım Modeli tabanlı Evrensel Arka Plan Modeli (GMM-UBM), çok sayıda

konuşmacıya ait konuşma sinyallerinden elde edilen akustik öznitelik vektörleri kullanılarak maksimum olabilirlik (maximum likelihood – ML) ilkesiyle ve beklentiyi maksimumlaştırma (Expectation Maximization – EM) algoritmasıyla eğitilmektedir [1]. Bu kapsamda, bireysel konuşmacı modelleri, UBM'den hareketle konuşmacının eğitim sinyalinden çıkarılan akustik öznitelik vektörleri aracılığıyla maksimum sonsal olasılık (maximum a Posteriori – MAP) adaptasyonu ile elde edilmektedir [23].

Süpervektör terimi, yüksek boyutlu vektörler aracılığıyla konuşmacı temsilini ifade ederken, bu vektörlerin kanal değişimi, mikrofon farkları ve ortam koşulları gibi etkenler altında performansının zayıf kaldığı çok sayıda çalışmada gösterilmiştir. Bu vektörlerin yüksek boyutlu yapısı, hem hesaplama hem de bellek açısından önemli bir yük oluşturmaktadır. Bu duruma çözüm olarak, eigen-channel [88], eigen-voice [87] ve Ortak Faktör Analizi (Joint Factor Analysis – JFA) [24] gibi boyut indirgeme ve kanal dengeleme teknikleri önerilmiştir. JFA'nın geliştirilmesi, daha etkili konuşmacı ayrımı sağlayan süpervektörlerin ortaya çıkmasını mümkün kılmıştır. Nihayetinde, konuşmacı ve kanal bilgisini aynı anda temsil eden, düşük boyutlu birleşik bir uzay tanımlayan ve ASV sistem performansını ciddi ölçüde artıran i-vektör yöntemi önerilmiştir [89]. Her ne kadar i-vektörler konuşmacı ve kanal bilgisini tam olarak ayıramasa da olasılıksal doğrusal ayırt edici analiz (PLDA) [90] gibi yöntemlerle birlikte kullanıldığında doğrulama başarımında önemli iyileşmeler sağlanmaktadır.

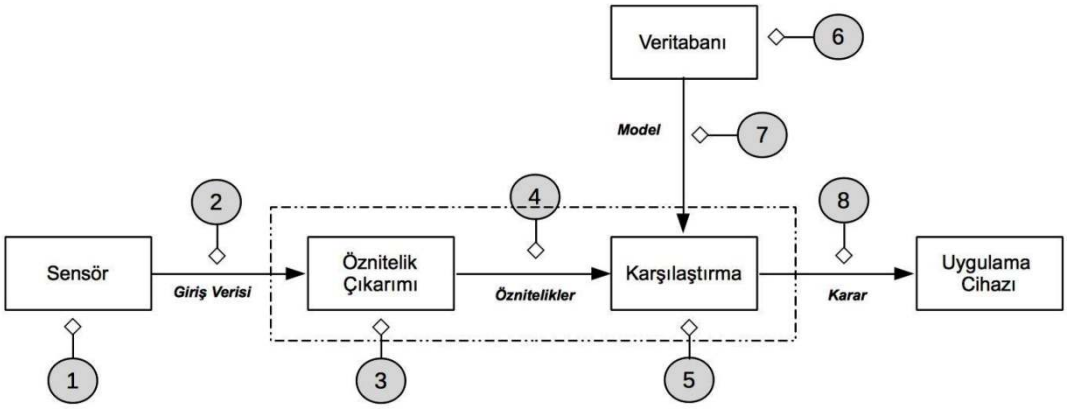
Derin öğrenme yöntemlerinin farklı alanlarda sağladığı başarı, ASV sistemlerinde de bu yaklaşımların kullanımını yaygınlaştırmıştır. Derin konuşmacı öznitelikleri (deep speaker embeddings), farklı uzunluklardaki ses sinyallerini sabit uzunlukta, konuşmacı kimliğine duyarlı vektörlere dönüştürmeyi amaçlamaktadır. Aynı konuşmacıya ait konuşma sinyallerinden elde edilen vektörlerin yakın, farklı konuşmacılardan elde edilenlerin ise uzakta konumlanması hedeflenmektedir. Bu fikir, kelimelerin vektörlerle ifade edildiği doğal dil işleme alanındaki word embedding yaklaşımıyla benzerlik göstermektedir. ASV sistemlerinde kullanılan konuşmacı öznitelikleri, denetimsiz (etiketsiz veriyle öğrenen) ve denetimli (etiketli veriyle ayırt edici şekilde eğitilen) olmak üzere iki gruba ayrılabilir. i-vektör yöntemi, üretici bir model üzerinden maksimum olabilirlik temelli eğitildiğinden denetimsiz bir temsildir. Buna karşın, x-vektör [83] gibi yöntemler, sınıflandırma kaybı kullanılarak eğitilen denetimli yaklaşımlar arasında yer almaktadır.

Literatürde derin öznitelik çıkarımı için farklı DNN mimarileri kullanılmıştır. Örneğin, d-vector [91] yaklaşımında akustik öznitelikler birden fazla tam bağlantılı (FC) katmandan geçirilerek işlenir ve her çerçeveden çıkan son gizli katman çıkışlarının ortalaması alınarak derin konuşmacı öznitelik vektörü oluşturulur. Ancak bu yöntem, özellikle büyük veri kümelerinde sınırlı performans sunmaktadır. Buna karşılık, zaman gecikmeli sinir ağı (TDNN) tabanlı x-vektör yönteminde [83, 92], çerçeve düzeyinde işlem yapılmakta olup ardından bir istatistiksel havuzlama katmanıyla bu çerçeve öznitelik vektörleri sabit uzunluklu derin öznitelik vektörlerine dönüştürülmektedir. Bu yapının gelişmiş hali olan ECAPA-TDNN [4], dört temel yenilik sunar: (i) kanal dikkat mekanizması ile konuşmacıya özel bilgilerin vurgulanması, (ii) farklı zaman çözünürlüklerinden bilgi yakalayabilen çok ölçekli öğrenim, (iii) çok seviyeli özniteliklerin birleştirilmesi ve (iv) öğrenmeyi kolaylaştıran artık bağlantılar. Bu gelişmeler, ECAPA-TDNN'nin günümüzde en başarılı derin öznitelik vektörü çıkarma yöntemlerinden biri olarak öne çıkmasını sağlamıştır. Bununla birlikte, daha verimli ve daha az parametreye sahip alternatif yapıların araştırılması sürmektedir. Bu doğrultuda, görüntü tanıma alanında geliştirilen Derin Artık Ağlar (ResNet) [93], konuşmacı tanıma uygulamalarında da etkin şekilde kullanılmaktadır. ResNet'in başlıca avantajı, residual learning yaklaşımı ile türev kaybı gibi derin ağ problemlerini azaltmasıdır. Literatürde yapılan çalışmalarda, örneğin [94]'de geleneksel havuzlama katmanlarının çıkarılmasıyla r-vector adını alan daha başarılı yapılar geliştirilmiştir. Bunun ardından, ResNet'in farklı varyasyonları da ASV problemi için uyarlanmıştır. Bu mimariler arasında Res2Net [95, 96], ResNext [95, 97], DF-ResNet [98, 99] ve ERes2Net [100] gibi çeşitli yapılar, doğruluk ve verimliliği artırma amacıyla önerilmiştir.

Son dönemde, üretken tabanlı yöntemler sınırlı veri senaryolarında performansı iyileştirmek için üretken adversaryal ağları (GAN) kullanmış ve sentetik veri üretimiyle daha etkili sistemler geliştirilmesini sağlamıştır. Günümüzde ise, özellikle Transformer tabanlı modellerin uzun bağlam bilgilerini yakalamada sağladığı avantajlarla, konuşmacı doğrulama sistemleri daha güçlü ve esnek bir hale gelmiştir. Bu kapsamda yapılan çalışmalar, alanın hem teorik hem de uygulamalı olarak sürekli gelişimine katkıda bulunmaya devam etmektedir.

2.2 Yanıltma Saldırısı Tespiti Üzerine Gerçekleştirilen Çalışmalar

CM sistemler, ASV sistemlerini hedef alan yanıltma saldırılarını tespit etmek ve engellemek amacıyla yıllar içinde birçok yenilikçi yaklaşımla geliştirilmiştir. Son yıllarda özellikle ses tabanlı konuşmacı tanıma sistemlerinde önemli ilerlemeler kaydedilmiş olsa da bu sistemler de diğer biyometrik tanıma teknolojileri gibi çeşitli saldırılara karşı savunmasızdır [87]. Şekil 2.1, genel biyometrik tanıma sisteminin yapısını ve bu sistemlerin maruz kalabileceği olası saldırı noktalarını göstermektedir. Biyometrik sistemlere yönelik saldırılar genel olarak iki ana grupta incelenebilir: doğrudan ve dolaylı saldırılar [87]. Doğrudan saldırılar, Şekil 2.1'de (1) numara ile gösterilen sensör düzeyinde gerçekleşir ve bu tür saldırılar, sistemin dış katmanında yürütüldüğü için genellikle şifreleme gibi dijital güvenlik önlemleriyle engellenemez. Literatürde yanıltma saldırıları (spoofing attacks) olarak adlandırılan [87] bu saldırılar, Uluslararası Standardizasyon Organizasyonu (ISO) tarafından sunum saldırısı (presentation attack) olarak da adlandırılmaktadır [101]. Doğrudan saldırılarda, sisteme yetkisi olmayan bir saldırgan, sahte yüz görüntüsü, yapay parmak izi veya sahte ses gibi manipüle edilmiş biyometrik veriler kullanarak başka bir kişiymiş gibi davranarak sisteme erişmeye çalışmaktadır. Öte yandan, dolaylı saldırılar sistemin iç işleyişini hedef alır ve genellikle siber saldırılar şeklinde ortaya çıkmaktadır. Örneğin, bir bilgisayar korsanı (hacker) öznitelik çıkarma veya karşılaştırma işlemlerini atlayarak (Şekil 2.1'de (3) ve (4) numaralı noktalar), veritabanındaki konuşmacı modellerini değiştirerek (Şekil 2.1'de (6) numaralı nokta) veya veri iletim sürecinde bulunan zayıf noktaları istismar ederek (Şekil 2.1'de (2), (4), (7) ve (8) numaralı noktalar) saldırı gerçekleştirebilir.



Şekil 2.1. Biyometrik bir sistemin genel yapısı muhtemel saldırı noktaları

Yanıltma saldırıları, ASV sistemlerinde önemli güvenlik açıkları yaratmakta olup, temelde fiziksel erişim (Physical Access – PA) ve mantıksal erişim (Logical Access – LA) olmak üzere iki grupta incelenmektedir. PA saldırılarında, genellikle kullanıcının daha önce kaydedilmiş konuşma kaydı yeniden çalınarak sistemin yanıltılması hedeflenirken, LA saldırıları TTS ve VC gibi yapay üretim teknikleriyle oluşturulmuş konuşma sinyallerinin kullanılmasını içermektedir [102]. Bu tehditlere karşı sistemlerin dayanıklılığını artırmak amacıyla geliştirilen karşı önlemler (CM), başlangıçta elle tasarlanmış öznitelikler ve geleneksel sınıflandırıcılar kullanılarak oluşturulmuş iken zamanla makine öğrenmesi ve derin öğrenme tabanlı yöntemlerle daha da gelişmiştir. Bu çabaların sistematik olarak değerlendirilmesini sağlayan ilk girişim, INTERSPEECH 2015 konferansında özel oturum olarak düzenlenen ASVspoof yarışması olmuştur [103]. Bu yarışmada, sistemlerin bilinen sahtecilik yöntemlerine karşı eğitilmesi ve daha önce karşılaşmadıkları yöntemlere karşı test edilmesi sağlanarak genelleme yetenekleri ölçülmüştür. Kullanılan veritabanı, hem gerçek hem de sahte konuşma sinyallerinden oluşmuş ve sahte sinyaller beş bilinen (S1–S5) ve beş bilinmeyen (S6–S10) yöntemle üretilmiştir. Bu sayede, sadece eğitim verisine aşina sistemlerin değil, aynı zamanda yeni saldırı türlerine karşı da etkili olan çözümlerin teşvik edilmesi amaçlanmıştır.

İlk yıllarda geliştirilen CM sistemleri, Mel Frekans Kepstrum Katsayıları (MFCC - Mel Frequency Cepstral Coefficients), Doğrusal Frekans Kepstrum Katsayıları (LFCC - Linear Frequency Cepstral Coefficients), Sabit-Q Kepstrum Katsayıları (CQCC - Constant Q Cepstral Coefficients) ve Modifiye Grup Gecikmesi (MGD - Modified Group Delay) gibi akustik özniteliklere dayalı olarak GMM (Gaussian Mixture Model) ve SVM (Support Vector Machine) sınıflandırıcılarıyla desteklenmiştir. Özellikle CQCC tabanlı yöntemler, 2015 ve 2017 yarışmalarında yüksek doğruluk oranlarıyla öne çıkmıştır [28]. 2017 yılında düzenlenen yarışmada odak, replay (tekrar oynatma) saldırılarının tespitine yönelmiş, veritabanı gerçek ortam koşullarını daha iyi yansıtacak biçimde yeniden düzenlenmiştir [104]. Saldırıların algılanmasında kullanılan yöntemler, önceki yarışmalara benzer şekilde geleneksel öznitelik çıkarımına dayansa da bazı çalışmalar farklı model çıktılarını birleştirerek (örneğin, ortalama alma ya da doğrusal regresyon) performans iyileştirmeyi hedeflemiştir [105–107]. 2019 yılında düzenlenen yarışmada ise saldırı türleri hem mantıksal hem fiziksel erişimi kapsayacak şekilde genişletilmiş ve bu doğrultuda LA ve PA senaryoları için

ayrı değerlendirme yapılmıştır [102]. Bu yarışmayla birlikte spektral ve faz temelli özniteliklerin kullanımı artarken, ResNet, CNN, LSTM ve LCNN gibi derin öğrenme mimarileriyle geliştirilen CM sistemleri öne çıkmıştır [108–112]. Ayrıca, AASIST [39] gibi spektral-zamansal dikkat mekanizmaları kullanan yeni modeller, geleneksel yöntemlere kıyasla daha yüksek doğruluk oranları elde ederek önemli performans iyileştirmeleri ortaya çıkarmıştır.

ASVspoof 2021 yarışmasında, önceki yıllardan farklı olarak doğrudan fiziksel ortamlarda kaydedilmiş tekrar oynatma saldırıları içeren daha gerçekçi bir veritabanı sunulmuştur [113]. Bu gelişme, fiziksel ortam değişkenliklerinin CM sistem performansı üzerindeki etkilerinin daha detaylı analiz edilmesini mümkün kılmıştır. Literatürde, bu dönemde geliştirilen sistemlerde klasik özniteliklerin (MFCC, LFCC, CQCC) yanı sıra enerji temelli ve spektral faz öznitelikleri gibi yenilikçi özellikler de yer almaya başlamıştır [114]. Derin öğrenme tabanlı yaklaşımların yaygınlaşmasıyla birlikte, ResNet ve Bi-LSTM gibi ağ yapıları tekrar oynatma saldırılarına karşı güçlü bir performans sergilemiştir [115-117]. Ayrıca, uçtan-uca yaklaşımlar dikkat çekmeye başlamış; ham dalga formu üzerinden çalışan CNN otokodlayıcıları, transformer tabanlı modeller ve spektrum analizine dayalı ağlar, öznitelik mühendisliğine ihtiyaç duymadan etkili sonuçlar vermiştir [118, 119]. 2021 yarışmasında, veri artırma teknikleri de yaygın bir şekilde uygulanmış, özellikle gürültü ekleme, kanal değişimi ve spektral bozulmalar gibi yöntemlerle sistemlerin genelleme yetenekleri güçlendirilmiştir [120, 121]. 2019 yılına kıyasla, 2021 yarışmasında kullanılan verilerin kapsamı genişlemiş, LA senaryosu için daha güncel TTS/VC teknikleriyle oluşturulmuş konuşmalar ve daha fazla konuşmacı yer almıştır. Bu farklar, 2021 veritabanını hem içerik hem de zorluk düzeyi açısından daha ileri bir noktaya taşımıştır.

2.3 Saldırıdan Haberdar Konuşmacı Doğrulama Üzerine Gerçekleştirilen Çalışmalar

[59] tarafından organize edilen SASV2022 yarışması CM ve ASV sistemlerinin entegrasyonunu teşvik etmeyi hedeflemiş ve bu yarışma kapsamında, katılımcılara yeni değerlendirme metrikleri ve protokoller sunulmuştur. En iyi sistemler %0,13'e kadar düşen hata oranlarıyla dikkat çekmiştir. [40] çalışmasında, SASV2022 yarışması kapsamında derin öznitelik vektörü (embedding) tabanlı birleştirme yöntemlerini

incelenmiş ve CatBoost gibi yöntemlerle SASV-EER metriklerinde önemli iyileştirmeler elde edilmiştir. Çalışmalar, doğrusal olmayan birleştirme yöntemlerinin skor birleştirme yaklaşımlarına kıyasla daha etkili olduğunu göstermiştir. [41] çalışmasında, çok görevli bir Conformer modeli geliştirilerek CM ve ASV görevleri birleştirilmiştir. ASVspoof 2019 LA veri kümesinde SASV-EER oranlarını %70 oranında iyileştiren bu model hem yerel hem de global bağlamları yakalamak için evrişimsel ve kendine dikkat mekanizmalarını bir arada kullanmıştır. [42] çalışması, SASV sistemlerinde skor seviyesinde birleştirme için olasılıksal bir yaklaşım önermiştir. İleri düzey çıkarım ve ince ayar stratejileri kullanarak SASV-EER oranlarını %19,31'den %1,53'e düşüren bu yöntem, ASV ve CM alt sistemlerinin entegrasyonunun performansı artırmada kritik olduğunu vurgulamıştır. [43] çalışmasında, CM ve ASV görevlerini birleştiren bir alt ağ (subnetwork) yaklaşımı geliştirilmiştir. ResNet yapısı üzerine eklenen küçük bir alt ağ eğiterek SASV2022 yarışmasının değerlendirme setinde %0,136 EER gibi etkileyici bir sonuç elde edilmiştir. Hafif tasarımı, yüksek doğruluğu korurken hesaplama maliyetlerini de en aza indirmiştir. [44], SASV2022 yarışması için temel sistemler sağlayarak basit skor toplama ve derin sinir ağı (DNN) tabanlı birleştirme stratejilerinin, bağımsız sistemlere kıyasla hata oranlarını önemli ölçüde düşürdüğünü göstermiştir. Çalışma, ASV ve CM alt sistemlerinin farklı seviyelerde entegre edilmesinin potansiyelini vurgulamıştır. [45], ASV ve CM sistemlerinin birlikte optimize edilmesinin tamamlayıcılığı artırdığını ortaya koymuş ve ek yardımcı veriler kullanarak hata oranlarında %27'lik göreceli bir azalma sağlamıştır. Ancak, sınırlı veriyle eğitimin ve aşırı öğrenmenin (overfitting) üstesinden gelmenin zorluklarına dikkat çekilmiştir. [46] ise, birden fazla CM sistemi (ResMax, AASIST) ve bir ASV sistemi (ECAPA-TDNN) kullanarak bir ensemble tabanlı yaklaşım geliştirmiştir. Derin öznitelik vektörlerinin birleştirilmesiyle SASV-EER değerini %5'e düşüren bu sistem, farklı modellerin bir araya getirilmesinin önemini vurgulamıştır. [47] ise, derin öznitelik vektörü ve skorların birleştirildiği çok seviyeli bir birleştirme çerçevesi sunmuş ve en iyi beş birleşim yöntemiyle SASV-EER değerini %0,89'a kadar düşürmüştür. Çalışma, tamamlayıcı bilgilerin en iyi şekilde kullanılmasını sağlamak için ensemble stratejilerinin gücünü ortaya koymuştur. [49] çalışması, SASV2022 yarışması için sahtekarlık farkındalığına sahip konuşmacı derin öznitelik vektörleri (SASE) geliştirmiş ve CM derin öznitelik vektörleri ile ASV derin öznitelik vektörlerini birleştirerek %0,28 SASV-EER değeri elde etmiştir. [64] çalışmasında, tek sınıflı

karışıklık kaybı ve rastgele derin öznitelik vektörleri örnekleme gibi yenilikçi yöntemleri kullanarak CM performansını artırmak amaçlanmıştır. SASV2022 yarışmasında seri sistemlerle %0,21 SASV-EER oranına ulaşarak temel sistemlerden çok daha iyi sonuçlar elde edilmiştir. [65] çalışması, tek sınıf kaybı (one-class loss) kullanarak hedef denemeler üzerinde odaklanan bir SASV entegrasyon ağı geliştirmiştir. Sistem, ASV ve CM alt sistemlerinden elde edilen derin öznitelik vektörlerinin doğrusal olarak birleştirilmesiyle, SASV2022 yarışmasında %0,14 gibi düşük hata oranlarına ulaşmıştır. [66], çok aşamalı bir eğitim çerçevesi kullanarak tek bir SASV derin öznitelik vektörü çıkarıcı model geliştirmiştir. Bu model, geniş ölçekli veriler ve yanıltma saldırısı tespiti için özel veri artırımı teknikleriyle, SASV2022 yarışmasında SASV-EER oranını %1,06'ya düşürmüştür. [67], çok katmanlı algılayıcı skor birleştirme modeli ve bütünleşmiş derin öznitelik vektörü projektörü gibi yenilikçi yaklaşımlarla SASV sistemlerini optimize etmiştir. Bu yöntemler, SASV2022 yarışmasında sırasıyla %0,56 ve %1,32 EER oranları elde etmiştir.

[48] çalışması, bağımsız bir CM modülüne gerek duymadan ASV sistemlerini derin öznitelik vektörü alanında sahtekarlık farkındalığına göre genelleştirmiştir. Bu yöntem, sinir ağı tabanlı arka uç sınıflandırıcı ile hata oranlarında %36,2'ye kadar iyileşme sağlamıştır. [50] çalışması, ASV ve CM görevleri için ortak özellik kullanımını ve Gauss arka uç birleşim yöntemini incelemiş, ASVspoof 2017 veri kümesinde başarılı sonuçlar elde etmiştir. Özellikle paralel birleştirme mimarisi, sahte ve hedef dışı denemeler arasındaki skor dağılımlarını daha iyi modelleyerek karar doğruluğunu artırmıştır. [51] çalışması, tekrarlı saldırı farkındalığına sahip bir ASV sistemi önererek, CM ve ASV sistemlerini birleştiren modüler bir arka uç yaklaşımı ile %21,77 EER iyileşmesi elde etmiştir. [52] çalışması, kontrastif kayıplı (contrastive loss) çok görevli öğrenme çerçevesiyle hem ASV hem de CM görevlerini ortak olarak ele alan bir sistem geliştirmiştir. Bu sistem, ASVspoof 2019'da %0,55'lik bir EER iyileştirmesi sağlayarak, ASV ve sahtekarlık tespiti için derin öznitelik vektörü tabanlı kararın avantajlarını göstermiştir. [53] çalışması, saldırı koşullarına dayanıklı ASV sistemleri için çok görevli bir öğrenme mimarisi geliştirmiş ve farklı saldırı türlerinde sistem performansını iyileştirmiştir. [54] çalışmasında, ASV ve CM alt sistemleri bağımsız olarak optimize edilerek elde edilen skorların norm-kısıtlı bir skor düzeyi birleştirme yöntemiyle birleştirilmesi ve bu sayede hedef kullanıcıları, hedef olmayan kullanıcıları ve yanıltma saldırılarını etkin bir şekilde ayırt etmeyi amaçlamaktadır. Bu

yaklaşım, SASV2022 yarışmasında %96'ya kadar iyileştirme sağlanmıştır. [55] çalışmasında ASV ve CM sistemlerini derin öznitelik vektörü seviyesinde birleştirerek, beklenen performans eğrisinin altında kalan alanı (the area under the expected - AUE) minimize etmeye dayalı yeni bir kayıp fonksiyonu kullanılmış ve yanıltma saldırılarına karşı %27'ye kadar hata oranı iyileşmesi elde edilmiştir. [56], ASV ve CM sistemlerini birlikte optimize etmek için takviye öğrenme (RL) tabanlı bir yöntem önermiştir. Bu yöntem, t-DCF metriğini optimize ederek, ASVSpoof 2019 veri kümesinde %20'lik bir iyileşme sağlamıştır. Çalışma, bu iki sistemin birbirine bağımlılığını dikkate alarak daha uyumlu bir performans elde edilmesini hedeflemiştir. [57]'de yazarlar, wav2vec 2.0 tabanlı özellik çıkartma ve kendi kendine ayrıştırma yöntemleriyle CM ve SASV sistemleri için daha etkili bir özellik temsili önererek, ASVspoof 2019 değerlendirme setinde %0,31 EER elde etmişlerdir. [58], çok görevli sınıflandırıcılar ve birden fazla kayıp fonksiyonunu kullanan uçtan uca bir SASV sistemi geliştirmiştir. Bu model hem ASV hem de CM görevlerini optimize ederek, ASVSpoof 2019 LA veri kümesinde SASV-EER metriğinde dikkate değer iyileştirmeler sağlamıştır. [62] çalışması, gözetimsiz alan uyarlaması (unsupervised domain adaptation) teknikleriyle ASV sistemlerini daha fazla sahtekarlık farkındalığına sahip hale getirme üzerine çalışmışlardır. Özellikle, ASVspoof 2019 veri setindeki konuşma sinyallerinden yararlanarak hem mantıksal hem de fiziksel erişim senaryolarında %36,1 ve %5,3 iyileştirme sağlamışlardır. [63] çalışması, birleştirilmiş ASV ve CM sistemlerinde, sinüs dalga filtrelerini ve spektral-zamansal grafik dikkat ağlarını kullanan bir SASV modeli önermiştir. Bu model, SASV2022 yarışmasının değerlendirme veri setinde SASV-EER oranını %6,73'ten %1,36'ya düşürmüştür.

2024 yılında düzenlenen ASVspoof 5 yarışması, SASV sistemlerini ve sahte ses tespiti (Deepfake Detection) yöntemlerini değerlendirmek amacıyla iki alt görev (track) üzerinden düzenlenmiştir. Track 2, SASV sistemlerini geliştirmek ve değerlendirmek için tasarlanmış olup katılımcıların ASV ve CM sistemlerini birleştirme becerilerini test etmektedir. [68] çalışması, SASV yarışması için CM modellerini ASV modelleriyle doğrusal birleştirme yöntemi kullanarak birleştirmiştir. ECAPA-TDNN modelini temel alarak ASV sistemini geliştiren ekip, veriyi genişletmek için çeşitli veri artırma teknikleri kullanmıştır. Önerilen sistem, kapalı koşulda 0,2814 min-aDCF ile dikkat çekici bir performans sergilemiştir. [69], Track 2'de ASV ve CM sistemlerini

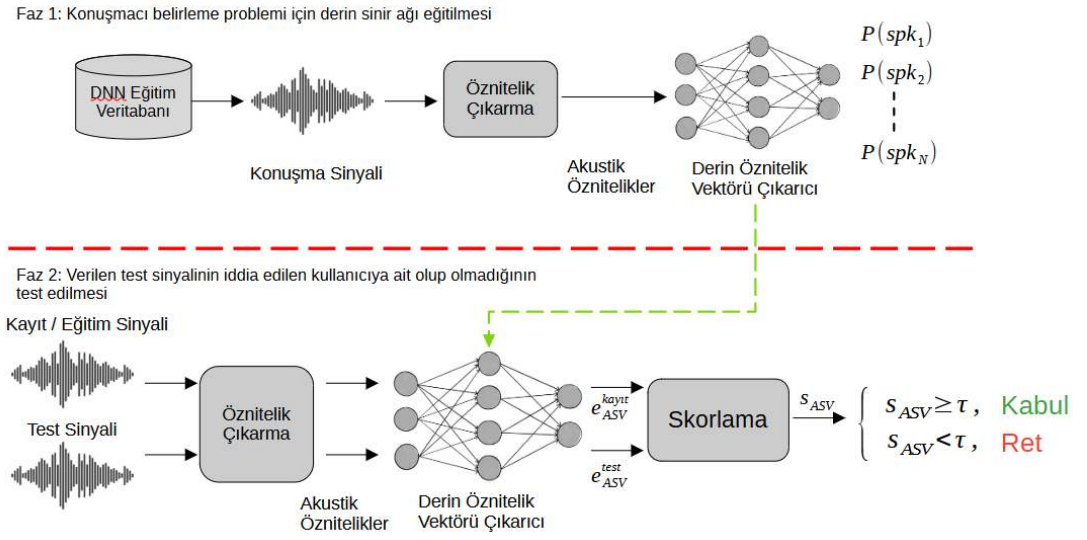
bir araya getiren bir kaskat yapı kullanmıştır. CM modeli olarak öz-denetimli bir öğrenme modeli olan Wav2Vec temel alınmıştır. Sonuçlar, açık koşullarda (open condition) 0,0756 min-a-DCF değerine ulaşarak yüksek bir performans göstermiştir. [70] çalışması, Track 2'de mel-spektrogram temelli giriş özniteliklerini kullanan ResNet tabanlı bir model geliştirmiştir. Model, farklı veri artırma teknikleriyle eğitilmiş ve elde edilen CM skorları ASV sistemiyle birleştirilmiştir. Kapalı koşullarda min-aDCF değerini 0,3173'e düşürmeyi başarmışlardır. [71], CM ve ASV modellerini birleştirmek için doğrusal olmayan birleştirme tekniklerini kullanmıştır. Geliştirilen sistemde, öz-denetimli öğrenen (SSL) modelleri ve ResNet tabanlı bir CM modeli entegre edilmiştir. Sonuçlar, Track 2 için 0,295 min-aDCF ile yüksek bir performans ortaya koymuştur. [72] çalışmasında, Track 2'de, Wav2Vec2 ve WavLM gibi kendine gözetimli modellerle CM skorları çıkarılmış ve TitaNet-Large modelini ASV skorlarını hesaplamak için kullanılmıştır. Bu skorlar, doğrusal olmayan birleştirme teknikleri ile birleştirilerek güçlü SASV bir performansı elde edilmiştir. [73], ASV ve CM modellerinin skorlarını entegre etmek için olasılık tabanlı bir yöntem önermiştir. ECAPA-TDNN ve ResNet tabanlı CM modelleri kullanarak Track 2'de %42'lik bir iyileşme sağlanmıştır. [74] çalışması, bağımsız CM ve ASV sistemlerini entegre etmek için bir puan ağırlıklandırma birleştirme yöntemi kullanmıştır. ASV için ResNet, CM için ise WavLM tabanlı modeller kullanan ekip, açık koşullarda 0,1203 min-aDCF ile dikkat çekici bir performans sergilemiştir. [75], Track 2'de CM ve ASV sistemlerinin log-likelihood skorlarını doğrusal olmayan bir birleştirme yöntemiyle birleştirmiştir. Bu yöntem, SASV metriklerinde önemli iyileşmelere yol açmıştır. [76], öz-denetimli öğrenme ve CNN tabanlı CM modellerini ASV modelleriyle doğrusal birleştirme yöntemiyle birleştirerek, açık koşulda 0,1156 min-aDCF ile başarılı bir performans sergilemiştir.

3. MATERYAL VE YÖNTEM

Bu bölümde tez çalışmasında kullanılan ASV modeller ve CM modellerin kısa bir özeti verilecektir. Daha sonra tezde kullanılan veri setleri tanıtılacak ve kullanılan değerlendirme metriklerinden bahsedilecektir. Son kısımda ise bu tez çalışması kapsamında önerilen yöntemler anlatılacaktır.

3.1 Otomatik Konuşmacı Doğrulama

Konuşmacı doğrulama sistemleri, bireylerin ses biyometrisini temel alarak kimlik tespiti yapan güvenlik teknolojileridir. Sistem, bir kişinin sisteme sunduğu konuşma sinyali ile önceden tanımlanmış bir kimliğe ait referans konuşma sinyalini karşılaştırarak, kullanıcının iddia ettiği kişi olup olmadığını değerlendirir. Test sırasında elde edilen konuşma sinyali (x^{tst}) bir mikrofon aracılığıyla toplanırken, sistemde önceden kayıtlı olan kimliğe karşılık gelen konuşma sinyali ise kayıt (enrollment) veya eğitim (training) sinyali (x^{enr}) olarak adlandırılır. Her iki sinyal, belirli bir ön işleme adımından geçirilerek akustik öznitelik vektörlerine dönüştürülür ve daha sonra eğitilmiş bir derin sinir ağı modeli yardımıyla sabit boyutlu derin öznitelik vektörü (embedding) vektörlerine dönüştürülür: $\mathbf{e}_{\text{asv}}^{\text{enr}}$ ve $\mathbf{e}_{\text{asv}}^{\text{tst}}$. Bu iki temsili vektör arasındaki benzerlik, sistemin karar vermesini sağlayacak sayısal bir skor üretir ve bu skor, tanımlanmış bir eşik değeri (τ_{asv}) ile karşılaştırılarak doğrulama işlemi gerçekleştirilmektedir.



Şekil 3.1: Derin Öznitelik Vektörleri Kullanılan Genel Bir ASV Sisteminin Yapısı.

Şekil 3.1'de gösterilen DNN tabanlı bir ASV sisteminin geliştirilme süreci genellikle iki ana fazdan oluşmaktadır. İlk faz, sınıflandırıcı modelin eğitildiği evredir. Bu aşamada, farklı konuşmacılardan toplanan ve her biri etiketlenmiş çok sayıda konuşma sinyali kullanılarak çok sınıflı bir sinir ağının eğitildiği aşamadır. Eğitim verisi $\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \dots, (x_M, y_M)\}$ biçiminde tanımlanmakta olup burada y_i, her bir örneğin hangi konuşmacıya ait olduğunu gösteren sınıf etiketleridir. Bu konuşma sinyalleri, öznitelik çıkarım fonksiyonu $\phi(\cdot)$ ile akustik öznitelik vektörlerine dönüştürülerek ($\mathbf{X}_i = \phi(x_i)$) eğitilecek sinir ağına giriş olarak uygulanır. Kullanılan öznitelikler genellikle MFCC, log-Mel spektrogram gibi zaman-frekans temsilleridir. Sinir ağının iç katmanları, bu özniteliklerden giderek daha soyut ve ayırt edici temsiller üretir. Son gizli katmandan çıkan derin öznitelik vektörleri, konuşmacıya özgü bilgileri içermekte olup bu vektörlerin doğruluğu, softmax çıkışı üzerinden tahmin edilen sınıf olasılıklarının gerçek etiketlerle karşılaştırılması ile ölçülür. Sinir ağının eğitimi amacıyla kullanılan kayıp fonksiyonu genellikle çapraz entropidir ve şu şekilde ifade edilmektedir:

$$\mathcal{L} = -\sum_{i=1}^{N} y_i \log \hat{y}_i \tag{3.1}$$

Burada y_i gerçek sınıf etiketini, $\hat{y}_i$ modelin tahmin ettiği olasılığı, N ise eğitimde kullanılan toplam örnek sayısını temsil etmektedir. Model parametreleri bu kaybı azaltacak şekilde optimize edilir ve eğitim süreci sonunda sinir ağı, konuşmacıları yüksek doğrulukla ayırt edebilen bir temsil öğrenmiş olur.

İkinci fazda, sınıflandırma işlevi değil, derin öznitelik vektörü (embedding) çıkarımı ön plana çıkmaktadır. Bu amaçla, eğitimde kullanılan DNN modelinin sınıflandırma katmanı çıkartılır ve geriye kalan yapı, sabit parametrelerle derin öznitelik vektörü çıkarma amacı ile kullanılmaktadır. Bu yapı $f(\Theta)$ ile gösterilir. Doğrulama işlemi sırasında öncelikle test (x^{tst}) ve kayıt (x^{enr}) konuşma sinyallerinden akustik öznitelik vektörleri $\mathbf{X}^{\text{enr}} = \phi(x^{\text{enr}})$, $\mathbf{X}^{\text{tst}} = \phi(x^{\text{tst}})$ elde edilir. Ardından:

$$\mathbf{e}_{\text{asv}}^{\text{enr}} = f(\mathbf{X}^{\text{enr}}; \Theta),\ \mathbf{e}_{\text{asv}}^{\text{tst}} = f(\mathbf{X}^{\text{tst}}; \Theta) \tag{3.2}$$

şeklinde iki sabit boyutlu derin öznitelik vektörü elde edilir. Bu vektörler arasındaki benzerlik, çoğunlukla kosinüs benzerliği ile ölçülür. Kosinüs benzerliği skoru

$$S = \frac{\mathbf{e}_{\text{asv}}^{\text{enr}^{\text{T}}} \cdot \mathbf{e}_{\text{asv}}^{\text{tst}}}{\|\mathbf{e}_{\text{asv}}^{\text{enr}}\| \cdot \|\mathbf{e}_{\text{asv}}^{\text{tst}}\|} \tag{3.3}$$

şeklinde hesaplanmaktadır. Elde edilen skor, eşik değeri τ_{asv} ile karşılaştırılarak son karar verilir. Eğer $S \geq \tau_{asv}$ ise test konuşmasının kayıtlı kişiyle eşleştiği kabul edilir; aksi durumda sistem kimlik iddiasını reddeder. Bu iki aşamalı yapı, güncel ASV sistemlerinde derin öğrenmenin gücünden faydalanarak hem ayırt edici hem de genellenebilir bir doğrulama performansı sunmayı amaçlamaktadır.

ASV sistemleri, özünde iki sınıflı bir karar mekanizması üzerine kuruludur ve bu nedenle sistemin verdiği kararlarda iki tür hata meydana gelebilir: Yanlış Kabul (False Acceptance – FA) ve Yanlış Reddetme (False Rejection veya Miss Detection – FR/Miss). Yanlış Kabul hatası, sistemin, aslında kimliği doğrulanan kişiye ait olmayan bir test konuşma sinyalini yanlışlıkla kabul etmesi durumunda oluşur. Diğer yandan, Yanlış Reddetme hatası ise, sistemin doğruluğundan emin olunabilecek bir test sinyalini hatalı şekilde reddetmesiyle ortaya çıkar. Bu iki hata türü dikkate alınarak ASV sistemlerinin genel doğruluk seviyesini ölçmek için en yaygın kullanılan performans kriterlerinden biri *Eşit Hata Oranı (Equal Error Rate – EER)*'dir. EER'yi hesaplayabilmek için öncelikle her iki hata türüne karşılık gelen oranların belirlenmesi gerekir.

Yanlış kabul oranı (False Acceptance Rate – FAR), sistemin kayıtlı olmayan bireyleri doğrulama ihtimalini gösterir. Güvenlik odaklı uygulamalar için bu oranın düşük tutulması esastır. FAR şu şekilde tanımlanır:

$$P_{\text{fa}}(\tau_{\text{asv}}) = \frac{S \geq \tau_{\text{asv}} \text{ olan yanlış sınamaların sayısı}}{\text{Toplam yanlış sınama sayısı}} \tag{3.4}$$

Yanlış ret oranı (False Rejection Rate / Miss Rate) ise sistemin kayıtlı kullanıcıları doğrulamada başarısız olma oranını temsil eder. Bu oranın düşük olması, sistemin kullanıcı dostu olduğunu gösterir:

$$P_{\text{miss}}(\tau_{\text{asv}}) = \frac{S < \tau_{\text{asv}} \text{ olan doğru sınamaların sayısı}}{\text{Toplam doğru sınama sayısı}} \tag{3.5}$$

Eşit Hata Oranı (EER), bu iki hata oranının birbirine eşit olduğu özel bir eşik ($\tau_{\text{asv}}^{\text{EER}}$) üzerinde tanımlanır. Bu noktadaki hata oranı:

$$\text{EER} = P_{\text{fa}}(\tau_{\text{asv}}^{\text{EER}}) = P_{\text{miss}}(\tau_{\text{asv}}^{\text{EER}}) \tag{3.6}$$

şeklinde ifade edilmektedir.

EER, sistemin genel denge performansını özetlemek açısından faydalı bir ölçüttür, ancak farklı uygulamalarda hata türlerinin etkileri aynı önemde olmayabilir. Örneğin güvenlik kritik sistemlerde yanlış kabuller daha büyük risk taşırken, kullanıcı dostu uygulamalarda yanlış reddetmeler daha olumsuz etkiler yaratabilir. Bu tür durumlar için daha esnek bir değerlendirme olanağı sunan *Sezim Bedel Fonksiyonu (Detection Cost Function – DCF)* performans metriği devreye girmektedir. DCF, özellikle ASV ve CM sistemlerinde kullanılan, sistemin performansını daha detaylı bir şekilde değerlendiren bir metriğin teorik bir uzantısıdır. DCF, bir sistemin doğruluk performansını, farklı hata türlerinin maliyetini dikkate alarak hesaplar. Bu metrik, sadece yanlış kabul oranı (FAR) ve yanlış reddetme oranı (FRR) gibi hataları ölçmekle kalmaz, aynı zamanda bu hataların ağırlıklarını (maliyetlerini) ve olayların önceden tahmin edilen olasılıklarını (prior probability) da hesaba katar.

ASV sisteminde ortaya çıkabilecek iki hata türü, uygulama bağlamına göre farklı maliyetlere sahip olabilir. Örneğin, güvenlik odaklı bir sistemde yanlış kabul çok daha kritik bir hata olarak kabul edilirken, müşteri hizmetlerinde yanlış reddetme daha büyük bir problem olabilir. DCF, bu farklı maliyetlerin sisteme olan etkisini değerlendirir.

DCF, bir ASV sistemi için False Acceptance Rate (FAR) ve False Rejection Rate (FRR) değerlerini şu şekilde birleştirir:

$$\text{DCF}\ (\tau_{\text{asv}}) = \ C_{\text{miss}}.P_{\text{miss}}(\tau_{\text{asv}}).P_{\text{tar}} + \ C_{\text{fa}}.P_{\text{fa}}(\tau_{\text{asv}}).(1 - P_{\text{tar}}) \qquad (3.7)$$

Denklem 3.7'de gösterilen:

- C_{miss}: Yanlış reddetmenin maliyeti (Cost of Miss).
- C_{fa}: Yanlış kabulün maliyeti (Cost of False Acceptance).
- $P_{\text{miss}}(\tau_{\text{asv}})$: Yanlış reddetme oranı (FRR).
- $P_{\text{fa}}(\tau_{\text{asv}})$: Yanlış kabul oranı (FAR).
- P_{tar}: Hedef konuşmacının önceden tahmin edilen olasılığı (Prior Probability)
- τ_{asv}: ASV sistem için tespit eşik değeri

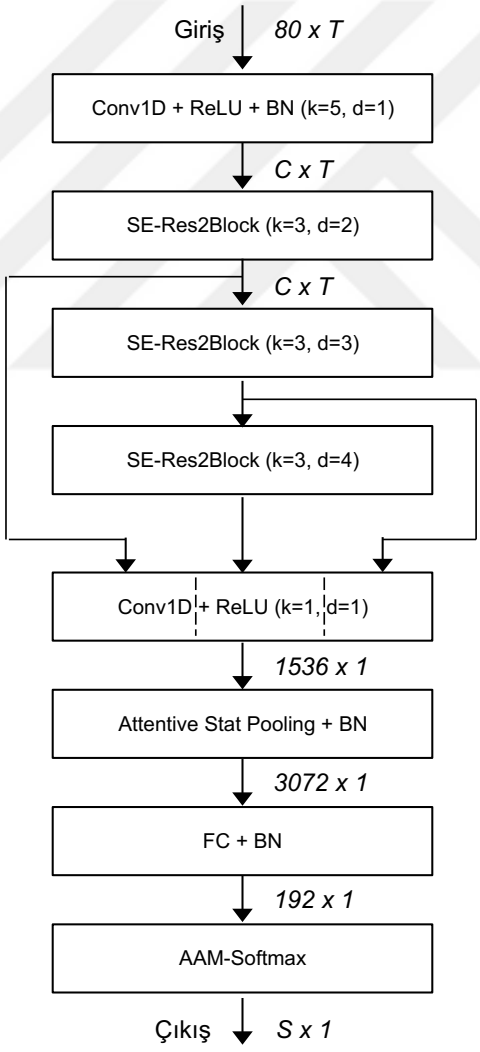
Bir sistemin performansını optimize etmek için DCF değerini en aza indiren eşik değeri belirlenebilir. Bu durumda, hesaplanan DCF değeri Minimum Detection Cost Function (MinDCF) olarak adlandırılır. MinDCF, sistemin karar eşik değerini

optimize ederek en düşük maliyetli performansı ölçmekte olup genellikle sistemlerin performansını karşılaştırmak için kullanılan bir metriktir.

3.2 Tez Çalışmalarında Kullanılan ASV Modeller

3.2.1 ECAPA-TDNN

Şekil 3.2'de gösterilen ECAPA-TDNN (Emphasized Channel Attention, Propagation, and Aggregation in TDNN) [4], konuşmacı doğrulama görevleri için geliştirilmiş bir Time Delay Neural Network (TDNN) tabanlı bir modeldir. Bu model, x-vektör [83] mimarisine dayalı olarak geliştirilmiş ve derin konuşmacı öznitelik vektörleri çıkartma performansını artırmak için çeşitli yenilikler sunmuştur.

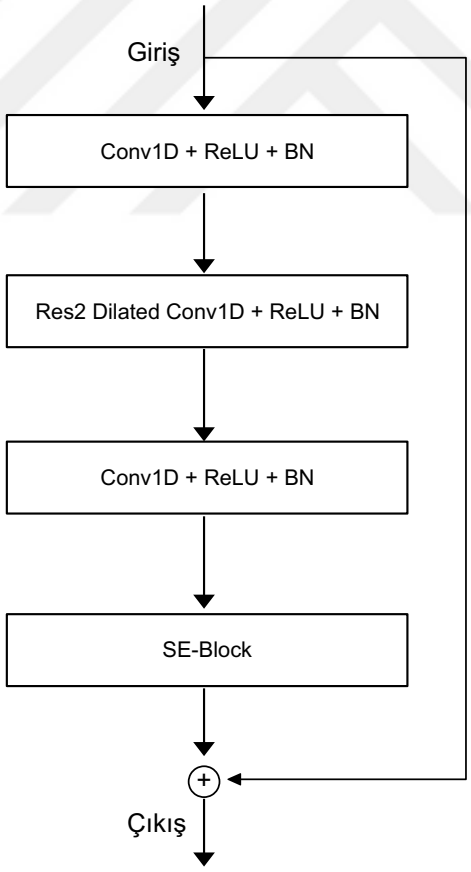


Şekil 3.2: ECAPA-TDNN mimarisi [4].

Modelin ilk temel yeniliği, *Kanal ve Bağlam Bağımlı İstatistik Havuzu* (Channel- and Context-Dependent Statistics Pooling) mekanizmasıdır. Bu mekanizma, zamansal

havuzlama sırasında kanal başına farklı dikkat mekanizmalarını kullanarak her bir kanal için çerçeve (frame) seviyesinde önem skorları hesaplamaktadır. Bu yaklaşım, konuşma kaydındaki global özellikleri (örneğin, ortam gürültüsü) dikkate alarak dikkat mekanizmasının tüm kayda uyum sağlamasına olanak tanımakta ve böylece modelin yalnızca önemli çerçevelere odaklanmasını sağlamaktadır.

İkinci yenilik, Şekil 3.3'te gösterilen bir boyutlu Squeeze-Excitation Res2Blocks yapı taşlarıdır. Bu bloklar, kanal bağımlı özelliklerin modelin öğrenmesine dahil edilmesini sağlayarak çok ölçekli (multi-scale) özelliklerin yakalanmasına olanak tanımaktadır. Sıkıştırma (squeeze) işlemiyle her kanalın ortalama vektörü hesaplandıktan sonra elde edilen vektörler, kanalları yeniden ölçeklendiren (excitation) ağırlıklarla birleştirilmektedir. Bu yöntem, ResNet tabanlı sistemlere benzer bir çalışma mantığına sahip olduğundan modelin parametre sayısını artırmadan performansı önemli ölçüde iyileştirmiştir.



Şekil 3.3: ECAPA-TDNN modelinde yer alan SE-Res2Block yapısı [4].

ECAPA-TDNN modelinde yer alan üçüncü yenilik, çok katmanlı öznitelik birleştirme (Multi-layer Feature Aggregation) ve toplama mekanizmalarıdır. Bu öznitelik, her bir

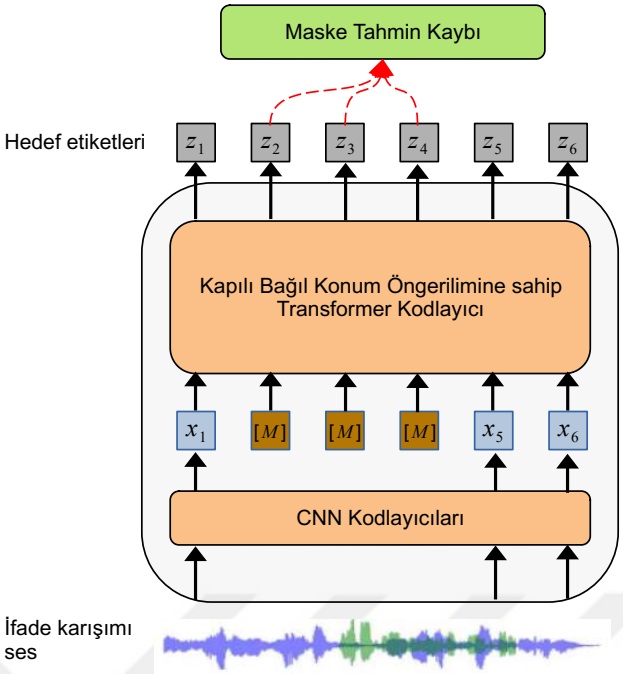
katmandan gelen bilgileri birleştirerek derin öznitelik vektörlerinin temsil kabiliyetinin daha yüksek hale gelmesini sağlamaktadır. Tüm SE-Res2Blocks'tan çıkan bilgilerin birleştirilmesi sayesinde daha zengin temsil gücü olan derin öznitelik vektörleri elde edilmektedir. Bu mekanizma, model parametre sayısını sınırlandırırken daha yüksek doğruluk sağlamaktadır.

ECAPA-TDNN modelinde derin konuşmacı öznitelik vektörlerini elde etmek amacıyla dikkat mekanizmasının uygulandığı istatistiksel havuzlama katmanı kullanılmaktadır. Bu yöntem, klasik istatistiksel havuzlama yerine kanal bazlı dikkat ağırlıkları aracılığıyla çerçeve seçimi yapmaktadır. Ayrıca, modelin çerçeve seviyesindeki bağlantı yapısını güçlendirmek için modelde ilave atlamalı bağlantılar yer almaktadır.

Bu model, özellikle VoxCeleb veritabanında mevcut TDNN tabanlı sistemlere kıyasla %19'a kadar daha düşük hata oranları ve %12,5 daha iyi MinDCF sonuçları elde ederek konuşmacı doğrulama sistemlerinde önemli bir ilerleme kaydetmiştir [4]. ECAPA-TDNN'nin yenilikçi yaklaşımı hem teorik hem de pratik olarak, TDNN tabanlı konuşmacı doğrulama mimarilerinin performansını önemli ölçüde iyileştirmiştir.

3.2.2 WavLM

WavLM (Waveform Language Model) [82], konuşma işleme görevleri için evrensel temsiller öğrenmeyi hedefleyen büyük ölçekli bir öz denetimli öğrenme (self-supervised learning – SSL) modelidir. Bu model, konuşma tanıma (automatic speech recognition - ASR) görevlerinin ötesine geçerek konuşmacı doğrulama, konuşmacı ayrıştırma ve konuşma ayrımı gibi çok yönlü konuşma işleme görevlerinde performansı artırmayı hedeflemektedir.



Şekil 3.4: WavLM model yapısı [82].

Şekil 3.4'te genel model yapısı gösterilen WavLM modelinin özellikleri arasında öncelikle WavLM çerçevesi öne çıkmaktadır. Bu çerçeve, maskeli konuşma tahmini ve gürültü giderme görevlerini birleştirerek konuşma içerik modellemesi yapmakla birlikte aynı zamanda çoklu konuşmacı ayrıştırma ve gürültülü konuşma senaryolarında da dayanıklılığı artırmaktadır. Gürültülü ve örtüşen konuşmaların maskelenmiş bölümlerinin orijinal etiketlerini tahmin etme yeteneği sayesinde, model konuşmacı kimliğinin tespit edilmesi ve konuşma iyileştirme performansını önemli ölçüde artırmaktadır. WavLM, HuBERT ve wav2vec 2.0 gibi modellerden farklı olarak, Transformer yapısında daha etkin bir konumlandırma mekanizması olan Gated Relative Position Bias özelliğini kullanan bir SSL modelidir. Bu mekanizma, konuşma içeriğine bağlı olarak göreli pozisyon önyargısını uyarlayacak şekilde optimize edilmiştir. Ayrıca model, LibriLight, GigaSpeech ve VoxPopuli gibi kaynaklardan toplanan toplam 94.000 saatlik çeşitli konuşma verileriyle eğitilmiştir. Bu geniş veri çeşitliliği sayesinde, yalnızca sesli kitap verilerine bağlı kalmayıp gerçek dünya koşullarında daha genelleştirilebilir hale gelmiştir.

Performans değerlendirmeleri göz önüne alındığında, WavLM modelinin SUPERB sıralamasında konuşmacı doğrulama, ASR ve konuşmacı ayrıştırma gibi birçok alt görevde HuBERT ve wav2vec 2.0 modellerinden daha iyi başarım gösterdiği

bilinmektedir. WavLM Large modelinin özellikle konuşmacı doğrulama ve konuşmacı ayrıştırma görevlerinde HuBERT Large modeline kıyasla daha yüksek performans sağladığı gösterilmiştir. VoxCeleb veri kümesi üzerinde yapılan deneylerde, WavLM Large modeli ECAPA-TDNN gibi bilinen sistemlerden daha düşük EER değerleri verdiği ve en iyi sistemlere göre yaklaşık %35'e varan göreceli hata oranı iyileştirmesi sağladığı görülmüştür. Ayrıca LibriCSS veri kümesi ile yapılan konuşma ayrıştırma deneylerinde, WavLM Large önceki modellerden %27,7 daha düşük kelime hata oranı elde edilmiştir.

WavLM modelinin katkıları arasında evrensel temsil öğrenimine yaptığı vurgu dikkat çekmektedir. Bu model yalnızca ASR için değil, aynı zamanda çoklu konuşmacı ayrıştırma ve gürültülü ortamlar gibi zorlu senaryolar için de genellenebilir temsiller öğrenmeyi hedeflemektedir. Maskeli konuşma denoising stratejisi, göreli pozisyonlama önyargısı ve geniş kapsamlı veri setleri gibi yenilikçi yaklaşımlar sayesinde model, hem içerik hem de konuşmacı temelli görevlerde üstün performans göstermektedir. WavLM, SUPERB ve diğer benchmark değerlendirmelerinde ASR ve konuşmacı doğrulama gibi görevlerde mevcut en iyi sistemleri geride bırakarak konuşma işleme alanında çok yönlü ve güçlü bir çözüm sunmaktadır.

Yukarıda belirtilen güçlü özellikleri ve üstün performansı nedeniyle WavLM yöntemi, bu tez çalışmasında kullanılan modellerden birisi olmuştur. WavLM modeli kullanılarak ASVspoof 5 veri setindeki konuşma sinyallerinden derin öznitelik vektörleri çıkarılarak SASV görevinde kullanılması sağlanmıştır. Böylece WavLM tabanlı özniteliklerin SASV görevindeki etkisi değerlendirilmiş ve modelin bu alanda sağlayabileceği katkılar araştırılmıştır.

3.2.3 TDNN

x-vektör modeli [83], ASV için kullanılan DNN tabanlı bir derin öznitelik vektörü çıkarım yöntemidir. Modelin temel amacı, değişken uzunluktaki konuşma parçalarını sabit boyutlu derin öznitelik vektörlerine dönüştürerek konuşmacı kimliği doğrulama süreçlerinde daha yüksek doğruluk oranları elde etmektir. Özellikle, i-vektör gibi geleneksel yöntemlere kıyasla daha güçlü bir performans sergilemekte olup konuşmacı özelliklerini etkili bir şekilde yakalamak için ileri düzey sinir ağı mimarilerinden yararlanmaktadır.

x-vektör sistemi, zaman gecikmeli sinir ağları (TDNN) ile çerçeve düzeyindeki akustik öznitelikleri işleyerek zaman boyutundaki bilgileri etkili bir şekilde modellemektedir. Çerçeve seviyesindeki özellikler istatistiksel havuzlama katmanı aracılığıyla sinyal seviyesine taşınmaktadır. İstatistiksel havuzlama katmanı, çerçeve düzeyindeki çıkışların ortalama ve standart sapmasını hesaplayarak sabit boyutlu derin öznitelik vektörleri oluşturularak konuşmacıya özgü bilgileri çıkarmaktadır. Sonraki tam bağlantılı katmanlar aracılığıyla, sinyal seviyesindeki bu derin öznitelik vektörleri işlenerek konuşmacı doğrulama için kullanılabilecek x-vektör derin öznitelik vektörleri elde edilir.

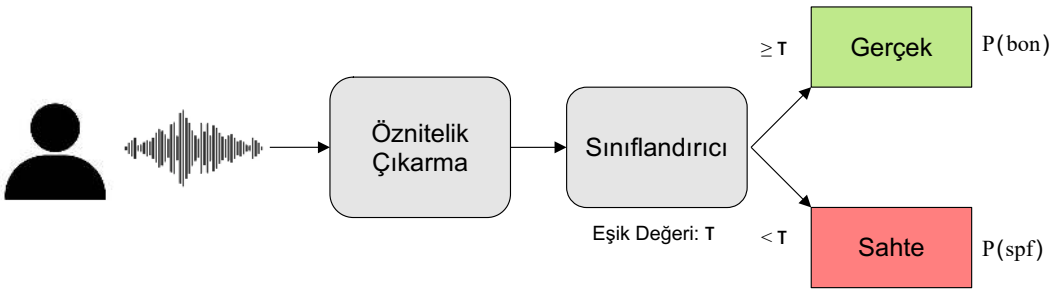
Model, konuşmacılar arasında ayrım yapmak için etiketli konuşma verileriyle eğitilir. Eğitim sırasında, her bir konuşma sinyali, konuşmacı etiketine dayalı olarak sınıflandırılır. Bu süreç, modelin konuşmacılar arasındaki farkları ayrıntılı şekilde öğrenmesini sağlar. Ayrıca, eğitim sürecinde kullanılan veri artırma teknikleri, modele gürültü ve yankı gibi çevresel değişimleri öğreterek modelin farklı senaryolarda dayanıklılığını artırmaktadır.

x-vektör sistemi, geleneksel i-vektor tabanlı sistemlere kıyasla daha üstün bir performans sergilemektedir. Gürültüye ve farklı akustik ortamlara karşı dayanıklılığı artırılmış olan model, PLDA (Probabilistic Linear Discriminant Analysis) gibi sınıflandırıcılarla birleştirildiğinde, konuşmacı doğrulama görevlerinde hata oranlarını önemli ölçüde düşürmektedir. Özellikle VoxCeleb ve NIST SRE16 gibi veri kümelerinde test edildiğinde, i-vektör yöntemlerine kıyasla %44'e kadar daha düşük hata oranları sağladığı gösterilmiştir [83].

3.3 Yanıltma Saldırısı Tespit Sistemi

Yanıltma saldırısı tespiti, ya da literatürde sıkça kullanılan terimiyle sahte konuşma (spoof speech) tanıma, otomatik konuşmacı doğrulama (ASV) görevine benzer şekilde ikili sınıflandırma problemi olarak ele alınmaktadır. Şekil 3.25'te gösterilen DNN tabanlı bir CM sisteminin genel yapısında amaç, bir konuşma sinyalinin gerçek (bonafide) ya da sahte (spoof) olup olmadığına karar verilmesidir. Derin sinir ağları (DNN) temelli bir CM sisteminin eğitimi, etiketli bir konuşma verisi kümesi kullanılarak gerçekleştirilir. Bu veri kümesi $\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \dots, (x_T, y_T)\}$ şeklindedir ve her bir x_i konuşma sinyali, buna karşılık gelen ikili sınıf etiketi $y_i \in$

$\{0, 1\}$ ile yer alır. Burada $y_i = 0$, gerçek bir konuşma sinyalini $y_i = 1$ ise sahte bir sinyali temsil etmektedir.



Şekil 3.5: DNN Tabanlı CM Sisteminin Genel Yapısı.

Her konuşma sinyali, eğitim sürecinde önce belirli bir ön işleme adımına tabi tutulur ve akustik öznitelik vektörleri $\mathbf{X}_i = \phi(x_i)$ elde edilir. Bu öznitelikler, konuşmanın zaman-frekans düzlemindeki temsilleridir ve literatürde en yaygın kullanılanları arasında LFCC (Linear Frequency Cepstral Coefficients), log-Mel filterbank çıktıları ve konuşma spektrogramları yer almaktadır. Elde edilen öznitelikler, uygun sınıflandırma etiketleri ile DNN modeline giriş olarak uygulanır. Eğitim süreci boyunca, sinir ağı bu giriş-çıkış ilişkilerini öğrenerek, sinyallerin sahte mi yoksa gerçek mi olduğunu ayırt edebilecek bir model haline gelmektedir.

Modelin çıkış katmanı, iki nörondan oluşur ve bu nöronlar sırasıyla gerçek ve sahte sınıflarını temsil eder. Her bir giriş için, softmax aktivasyon fonksiyonu yardımıyla her sınıfa ait olasılık değerleri hesaplanır:

$$P(\text{bon}) = \frac{e^{z_1}}{\sum_{j=1}^{2} e^{z_j}}\ ,\ P(\text{spf}) = \frac{e^{z_2}}{\sum_{j=1}^{2} e^{z_j}} \tag{3.8}$$

Burada z_1 ve z_2, sırasıyla gerçek ve sahte sınıflarına karşılık gelen nöronların aktivasyon skorlarıdır. Elde edilen bu olasılıklar, genellikle karar skoru oluşturmak amacıyla logaritmik olabilirlik oranı (log-likelihood ratio) biçiminde birleştirilir:

$$S_{cm} = \log P(\text{bon}) \ - \log P(\text{spf}) \tag{3.9}$$

Bu skor, belirli bir eşik değeri (τ_{cm}) ile karşılaştırılarak karar verme işlemi gerçekleştirilir. Eğer $S_{cm} \geq \tau_{cm}$ ise sistem sinyali gerçek olarak kabul eder ($\hat{y} = 1$); aksi durumda sahte olarak sınıflandırır ($\hat{y} = 0$).

Modelin eğitimi sırasında genellikle ikili çapraz entropi (binary cross-entropy) kayıp fonksiyonu kullanılmaktadır. Bu fonksiyon, modelin tahmin ettiği sınıf olasılıkları ile gerçek sınıf etiketleri arasındaki farkı minimize etmeyi amaçlar ve

$$\mathcal{L} = -\frac{1}{T}\sum_{i=1}^{T} y_i \cdot \log \hat{y}_i + (1 - y_i) \cdot \log(1 - \hat{y}_i) \tag{3.10}$$

şeklinde tanımlanmaktadır. Kayıp fonksiyonunun minimum yapılması için yaygın olarak kullanılan optimizasyon teknikleri arasında stokastik türev azaltma (stochastic gradient descent - SGD) ve Adam algoritmaları yer almaktadır.

CM sistemlerinde, tıpkı ASV görevinde olduğu gibi, iki tür sınama gerçekleştirilir: gerçek (bonafide) ve sahte (spoof) konuşma sinyalleri. Gerçek sınamalar, orjinal konuşma sinyallerini içerirken, sahte sınamalar; TTS, VC veya tekrar oynatma gibi tekniklerle üretilmiş, güvenilir olmayan konuşma sinyallerini temsil etmektedir. Bu sınamalara verilen sistem yanıtı doğrultusunda *Yanlış Ret (Miss Detection – Miss)* ve *Yanlış Kabul (False Acceptance – FA)* veya diğer bir ifadeyle *Yanlış Alarm (False Alarm)* hataları olmak üzere iki tür hata ortaya çıkmaktadır. Yanlış ret, sistemin gerçek bir konuşma sinyalini sahte olarak sınıflandırması; yanlış kabul ise sahte bir sinyali yanlışlıkla gerçek olarak değerlendirmesi durumudur. Bu hata türlerine karşılık gelen oranlar, belirli bir karar eşiği τ_{cm} kullanılarak aşağıdaki şekilde hesaplanır:

$$P_{\text{fa}}^{\text{cm}}(\tau_{\text{cm}}) = \frac{S_{\text{cm}} \geq \tau_{\text{cm}} \text{ olan yanlış sınamaların sayısı}}{\text{Toplam yanlış sınama sayısı}} \tag{3.11}$$

ve

$$P_{\text{miss}}^{\text{cm}}(\tau_{\text{cm}}) = \frac{S_{\text{cm}} < \tau_{\text{cm}} \text{ olan gerçek sınamaların sayısı}}{\text{Toplam gerçek sınama sayısı}} \tag{3.12}$$

Bu iki oran eşit hale geldiğinde elde edilen eşik seviyesi ($\tau_{\text{cm}}^{\text{EER}}$) CM sistemlerinin performansını değerlendirmek için sıkça kullanılan *Eşit Hata Oranı (Equal Error Rate – EER)* metriğinin temelini oluşturmaktadır. EER, sahte ve gerçek konuşmaların sistem tarafından ne ölçüde doğru ayrılabildiğini gösteren bir metrik olmakla birlikte

$$\text{EER}_{\text{cm}} = P_{\text{fa}}(\tau_{\text{cm}}^{\text{EER}}) = P_{\text{miss}}(\tau_{\text{cm}}^{\text{EER}}) \tag{3.13}$$

şeklinde tanımlanır. Ancak yanıltma saldırısı tespit sistemleri, çoğu zaman ASV sistemleriyle birlikte çalıştığı için, sadece CM hatalarına odaklanan EER metriği,

sistemin genel kullanım senaryosunu tam olarak yansıtmamaktadır. Bu nedenle, CM ve ASV sistemlerinin birlikte çalıştığı durumlardaki genel performansı değerlendirmek amacıyla *tandem Detection Cost Function (t-DCF)* metriği geliştirilmiştir [124]. t-DCF metriği hem saldırı tespiti hem de ASV doğrulamasının hata ve maliyetlerini dikkate alan bir metrik olup, uygulama koşullarına göre hangi hata türlerinin daha maliyetli olduğuna dair özelleştirilebilir bir değerlendirme sunmaktadır.

t-DCF, CM ve ASV sistemlerine ait belirli bir eşik değeri çifti ($(\tau_{\text{asv}}, \tau_{\text{cm}})$) üzerinden, dört farklı hata türünü ve bunlara karşılık gelen önsel olasılıkları içeren aşağıdaki formülle tanımlanır [124]:

$$\text{t} - \text{DCF}(\tau_{\text{asv}}, \tau_{\text{cm}}) = C_{\text{miss}}^{\text{asv}} \cdot \pi_{\text{tar}} \cdot \left(1 - P_{\text{miss}}^{\text{cm}}(\tau_{\text{cm}})\right) \cdot P_{\text{miss}}^{\text{asv}}(\tau_{\text{asv}}) + C_{\text{fa}}^{\text{asv}} \cdot \pi_{\text{non}} \cdot \left(1 - P_{\text{miss}}^{\text{cm}}(\tau_{\text{cm}})\right) \cdot P_{\text{fa}}^{\text{asv}}(\tau_{\text{asv}}) + C_{\text{fa}}^{\text{cm}} \cdot \pi_{\text{spf}} \cdot P_{\text{fa}}^{\text{cm}}(\tau_{\text{cm}}) \cdot \left(1 - P_{\text{miss,spf}}^{\text{asv}}(\tau_{\text{asv}})\right) + C_{\text{miss}}^{\text{cm}} \cdot \pi_{\text{tar}} \cdot P_{\text{miss}}^{\text{cm}}(\tau_{\text{cm}}) \quad (3.14)$$

Denklem 3.14'te:

- $C_{\text{miss}}^{\text{asv}}$ ve $C_{\text{fa}}^{\text{asv}}$, ASV sisteminin yanlış reddetme ve yanlış kabul etme durumlarında oluşan maliyetleri;
- $C_{\text{fa}}^{\text{cm}}$ ve $C_{\text{miss}}^{\text{cm}}$, CM sistemine ait benzer hata maliyetlerini;
- π_{tar}, π_{non} ve π_{spf} ise sırasıyla doğru kullanıcı (target), farklı kişi (nontarget) ve sahte (spoof) denemelerine ait önsel olasılıkları temsil etmektedir.

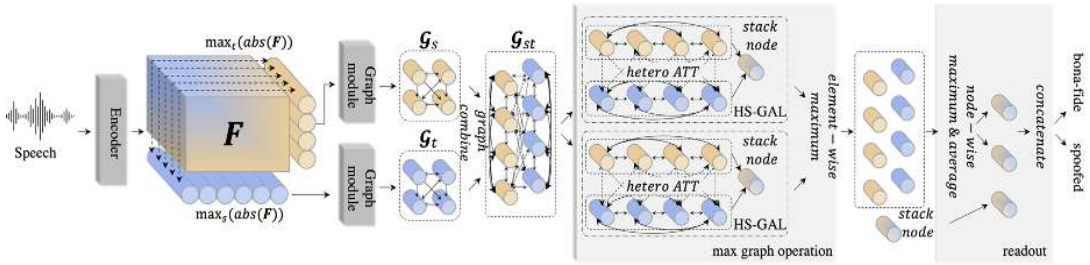
Denklem 3.14 incelendiğinde, kusursuz bir CM sistemi olması durumunda yani $P_{\text{fa}}^{\text{cm}} = P_{\text{miss}}^{\text{cm}} = 0$ olduğunda, t-DCF metriği yalnızca ASV sisteminin hata maliyetlerini kapsayan klasik DCF ifadesine indirgenmektedir. Benzer şekilde, mükemmel bir ASV sistemi olması durumunda, yani $P_{\text{fa}}^{\text{asv}} = P_{\text{miss}}^{\text{asv}} = 0$ ise t-DCF yalnızca CM sistemine özgü maliyetleri içeren bir değerlendirmeye dönüşmektedir.

3.4 Tez Çalışmalarında Kullanılan CM Modeller

3.4.1 AASIST

AASIST (Audio Anti-Spoofing Using Integrated Spectro-Temporal Graph Attention Networks) [39], sahte ses tespit görevlerinde etkili ve verimli bir uçtan uca çözüm sunmak için tasarlanmış bir modeldir. Model, şekil 3.6'da görüldüğü gibi hem spektral

hem de zamansal alanlarda sahtecilik belirtilerini tespit etmek için heterojen bir graf dikkat katmanı (HS-GAL) ve yeni bir maksimum grafik operasyonu (Max Graph Operation - MGO) içeren yenilikçi bir mimariye sahiptir.



Şekil 3.6: AASIST modelinin genel yapısı [39].

AASIST'in temel yapı taşlarından biri, spektral ve zamansal alanları temsil eden iki farklı grafın birleştirilmesi ve bu grafların heterojen bir dikkat mekanizmasıyla işlenmesidir. Spektral ve zamansal graf düğümleri, farklı özellik uzaylarında bulunmakta olup HS-GAL katmanında bu düğümler arasındaki ilişkiler üç farklı projeksiyon vektörü ile ortaya çıkarılmaktadır. HS-GAL ayrıca bir "stack node" adı verilen özel bir düğüm yer almaktadır. Bu düğüm, spektral ve zamansal bilgiler arasındaki ilişkileri toplamak için tasarlanmış olup diğer düğümlere bilgi iletmeden yalnızca veri toplama görevi gerçekleştirmektedir. Bu özellik, modelin heterojen verilerden etkili bir şekilde bilgi çıkarmasını sağlamaktadır.

AASIST modelinin bir diğer önemli bileşeni, MGO adı verilen bir mekanizmadır. MGO, paralel iki grafik dalının kullanıldığı ve her dalda iki HS-GAL katmanı ile grafik havuzlama işlemleri gerçekleştirildiği bir mekanizmadır. Bu iki dalın çıktıları, maksimum eleman seçimi ile birleştirilir ve bu işlem, farklı sahtekarlık belirtilerinin aynı anda tespit edilmesini sağlamaktadır. Ayrıca, MGO'nun çıktılarını birleştirmek için özel bir "readout" mekanizması yer almaktadır. Bu mekanizma, düğüm seviyesinde maksimum ve ortalama değerleri hesaplar ve stack node ile birleştirerek nihai bir çıktı katmanı oluşturur.

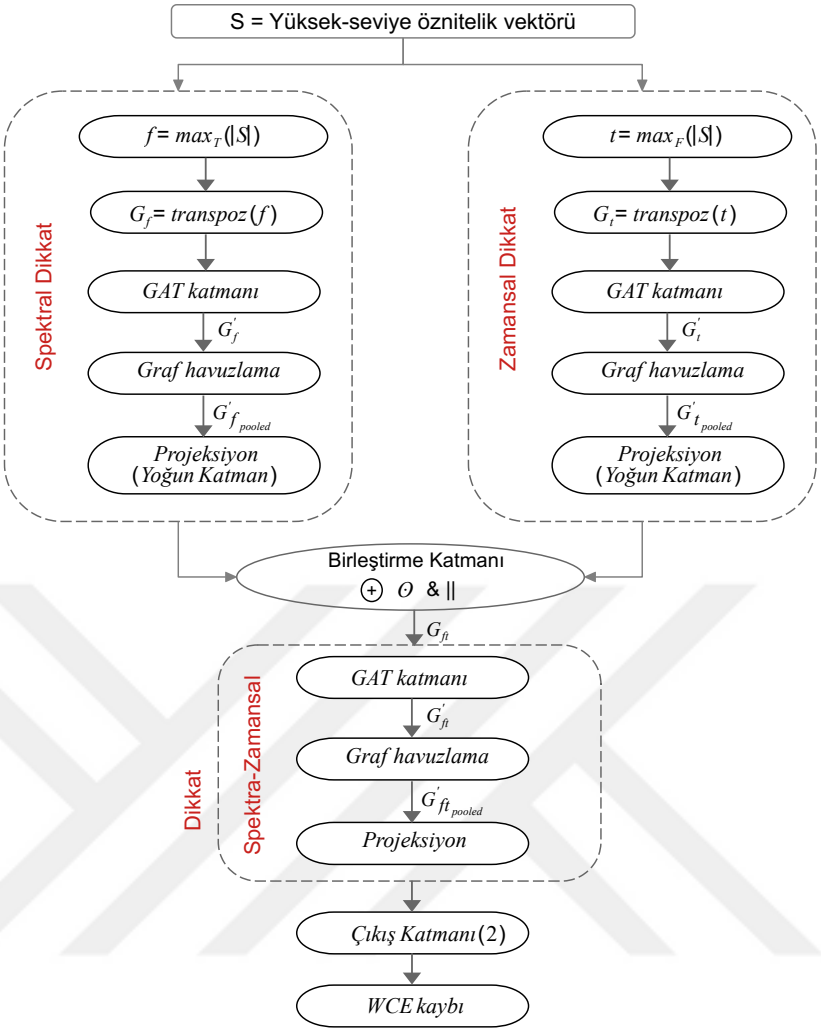
AASIST, RawNet2 tabanlı bir kodlayıcı (encoder) kullanarak ham dalga formundan yüksek seviyeli özellikler çıkaran bir modeldir. Bu kodlayıcı, sinyalin spektral ve zamansal bileşenlerini öğrenmek için Sinc-convolution katmanı ve artık bloklar (residual blocks) içermektedir. Bu yapı, modelin doğrudan ham konuşma verileriyle çalışmasını sağladığından ön işlem ihtiyacını ortadan kaldırmaktadır.

Model, ASVspoof 2019 veri seti üzerinde test edilmiş ve hem min-tDCF hem de EER metriklerinde önceki yöntemlere kıyasla %20'ye varan iyileştirmeler sağlamıştır [39]. AASIST, yalnızca 297 bin parametre ile oldukça verimli bir model olup daha az parametreli bir versiyon olan AASIST-L modeli, yalnızca 85 bin parametreyle çoğu rakip sistemi geride bırakmıştır. AASIST, sahtecilik tespiti görevlerinde en iyi performans gösteren sistemlerden biri olarak dikkat çekmiştir ve hem spektral hem de zamansal alanlardaki bilgileri entegre etme yeteneğiyle ön plana çıkmaktadır.

3.4.2 RawGAT-ST

RawGAT-ST (Raw Graph Attention Network with Spectro-Temporal attention) [84], konuşma doğrulama, sahtekarlık tespiti ve derin sahte (deepfake) ses tespiti gibi görevler için geliştirilen uçtan uca bir modeldir. Model, ham dalga formlarından spektral ve zamansal özellikler öğrenerek, bu özellikleri grafik tabanlı bir dikkat mekanizmasıyla işlemektedir. Bu sayede, sahte ve gerçek sesler arasındaki ayırt edici bilgileri spektral ve zamansal alanlarda aynı anda ortaya çıkarmaktadır.

RawGAT-ST modelinde de AASIST modelinde olduğu gibi ham konuşma sinyalleri üzerinden işlem yapıldığından geleneksel öznitelik çıkarma adımlarına ihtiyaç duyulmamaktadır. Modelin ön cephesi, ham dalga formunu Sinc konvolüsyon katmanları aracılığıyla işleyerek zaman-frekans bileşenlerini çıkarmak üzere tasarlanmıştır. Bu katman, ham ses verisinde band geçiren filtreler olarak çalışan parametrik bir yapı sağlamaktadır. Bu süreç, modelin hem frekans hem de zaman alanında anlamlı temsiller oluşturmasını kolaylaştırmaktadır.



Şekil 3.7: RawGAT-ST modelinin genel yapısı [84]

Şekil 3.7'de genel yapısı gösterilen RawGAT-ST modeli, spektral ve zamansal dikkat bloklarından oluşan bir grafik dikkat ağı yapısına sahiptir. Spektral dikkat blokları, frekans alt bantlarındaki ayrımcı bilgileri yakalarken, zamansal dikkat blokları zaman aralıklarındaki kritik bilgilere odaklanmaktadır. Bu iki bloktan elde edilen grafikler daha sonra bir spektral-zamansal dikkat bloğunda birleştirilir. Bu blok, spektral ve zamansal bilgiler arasındaki ilişkileri modelleyerek sahtecilik tespitinde performansta iyileştirme sağlamaktadır.

Grafik havuzlama mekanizması, grafik boyutlarını küçülterek en bilgilendirici düğümleri seçmek için kullanılan bir yapıdır. Bu işlem, grafikteki gürültüyü azaltır ve modelin daha ayrımcı özelliklere odaklanmasını sağlamaktadır. Grafik havuzlama sürecinde öğrenilebilir projeksiyon vektörleri aracılığıyla düğüm özellikleri sıralanır ve yalnızca en önemli düğümler seçilmektedir.

Model seviyesinde birleştirme, spektral ve zamansal grafiklerin çıkışlarının birleştirilmesini içerir. Birleştirme işlemi, öğe bazlı toplama, çarpma veya birleştirme (concatenation) yöntemleriyle gerçekleştirilir. Bu süreç, spektral ve zamansal dikkat mekanizmalarından gelen tamamlayıcı bilgilerin etkili bir şekilde birleştirilmesini sağlar. Sonuç olarak, model iki sınıf arasında (gerçek veya sahte) bir tahmin yapmak için projeksiyon ve çıkış katmanları kullanır.

ASVspoof 2019 LA veri kümesinde test edilen RawGAT-ST, %1,06 gibi düşük bir EER ve 0,0335 min-tDCF ile bugüne kadar rapor edilen en iyi sonuçlardan birini elde etmiştir. Ablasyon çalışmaları, spektral dikkat mekanizmasının zamansal dikkate göre daha kritik olduğunu, ancak her ikisinin de birlikte kullanımının performansı önemli ölçüde artırdığını göstermiştir. Grafik havuzlama mekanizmasının kullanımı, modelin ayrım gücünü daha da artırarak genel başarımı sağlamlaştırmıştır.

3.5 Kullanılan Veri Setleri

3.5.1 VoxCeleb2

VoxCeleb2 [125] veri seti, geniş kapsamlı konuşmacı doğrulama ve tanıma görevleri için kullanılan büyük ölçekli bir veri seti olup YouTube gibi çevrimiçi platformlardan otomatik olarak toplanmış ses ve video kayıtlarından oluşmaktadır. Veri seti, doğal konuşma örnekleriyle birlikte çeşitli akustik ortamlarda ve gürültü koşullarında kaydedilmiş verilerden meydana gelmektedir. Metin bağımsız bir yapıya sahip olan VoxCeleb2, farklı aksanlara, yaş gruplarına ve cinsiyetlere sahip konuşmacıları içerdiğinden ASV modellerinin gerçek dünya koşullarında genelleştirilebilirliğine olanak sağlamaktadır. Çizelge 1'de bu veri setine ait bilgiler verilmiştir.

Çizelge 1: VoxCeleb2 veri setine ilişkin bilgiler.

	Geliştirme	Test
Konuşmacı sayısı	5994	118
Video sayısı	145569	4911
Konuşma kaydı sayısı	1092009	36237

Geliştirme setinde 5.994 konuşmacı, 145.569 video ve 1.092.009 konuşma sinyali bulunurken, test seti 118 konuşmacı, 4.911 video ve 36.237 konuşma sinyali içermektedir.

3.5.2 ASVspoof 2019

ASVspoof 2019 [13] veri seti, ASV sistemlerinin sahtekarlık saldırılarına karşı dayanıklılığını değerlendirmek amacıyla oluşturulmuş kapsamlı bir veri tabanıdır. Bu veri seti, iki ana senaryo altında üç temel sahtekarlık türünü içerir:

1. *Mantıksal Erişim (LA) Senaryosu:*
   - *Saldırı Türleri:* Sentetik konuşma ve ses dönüştürme saldırıları.
   - *Veri Kaynağı:* VCTK veri tabanından seçilen 107 konuşmacının (46 erkek, 61 kadın) kayıtları kullanılmıştır.
   - *Sahte Kayıtlar:* Gelişmiş konuşma sentezi ve ses dönüştürme teknikleri kullanılarak oluşturulmuştur.
   - *Veri Dağılımı:*
     - *Eğitim Seti:* 2.580 gerçek, 22.800 sahte kayıt.
     - *Geliştirme Seti:* 2.548 gerçek, 22.296 sahte kayıt.
     - *Değerlendirme Seti:* 7.355 gerçek, 63.882 sahte kayıt.
2. *Fiziksel Erişim (PA) Senaryosu:*
   - *Saldırı Türleri:* Tekrar oynatma (replay) saldırıları.
   - *Veri Kaynağı:* VCTK veri tabanından seçilen 20 konuşmacının (10 erkek, 10 kadın) kayıtları kullanılmıştır.
   - *Sahte Kayıtlar:* Farklı oynatma ve kayıt cihazları, ortamlar ve mesafeler kullanılarak oluşturulmuştur.
   - *Veri Dağılımı:*
     - *Eğitim Seti:* 5.400 gerçek, 48.600 sahte kayıt.
     - *Geliştirme Seti:* 5.400 gerçek, 24.300 sahte kayıt.
     - *Değerlendirme Seti:* 18.090 gerçek, 153.054 sahte kayıt.

3.5.3 ASVspoof 5

ASVspoof 5 [85], sahtekarlık saldırıları ve derin sahte (deepfake) konuşma tespiti üzerine çalışan otomatik konuşmacı doğrulama (ASV) sistemlerinin performansını değerlendirmek amacıyla geliştirilmiş bir veri setidir. Bu veri seti, önceki ASVspoof sürümlerinden önemli farklılıklar gösterir; özellikle crowdsourced yöntemlerle geniş bir konuşmacı topluluğu ve çeşitli akustik koşullar göz önünde bulundurularak oluşturulmuştur. Veri seti, Mantıksal Erişim (LA) ve derin sahte (DF) senaryolarını birleştirirken, CM ve SASV alt kategorileri olmak üzere iki farklı yarışma alt kategorisine (track) odaklanmaktadır. Çizelge 2'de detaylı olarak verilen veri seti bölümü, konuşmacı sayısı, gerçek ve sahte örnek sayısı gibi istatistiksel bilgiler, bu veri setinin kapsamı ve çeşitliliği hakkında genel bir bakış sunmaktadır.

Çizelge 2: ASVspoof 5 veri setine ilişkin bilgiler.

Veriseti Bölümü	Konuşmacı Sayısı (Kadın / Erkek)	Gerçek (Bonafide)	Sahte (Spoof)
Eğitim	196 / 204	18797	163560
Geliştirme	392 (196) / 393 (202)	31334	109616
Değerlendirme T1	370 / 367	138688	542086
Değerlendirme T2	370 (194) / 367 (173)	100708	395924

ASVspoof 5, TTS ve VC yöntemleri ile üretilmiş sahte sinyaller içermektedir. Bu sinyaller, yeni nesil derin öğrenme tabanlı modeller (örneğin, GlowTTS, Tacotron2, ZMM-TTS gibi) ve adversaryal saldırılar (Malafide ve Malacopula filtreleri) kullanılarak oluşturulmuş olup toplamda 32 farklı yöntemle üretilmiş sahte sesler bulunmaktadır.

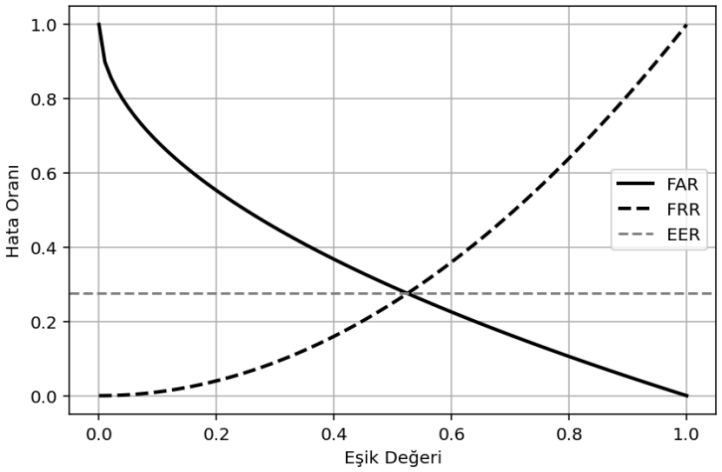
3.6 Değerlendirme Kriterleri

3.6.1 Equal error rate (EER)

EER, ikili sınıflandırma problemlerinde sistem performansını değerlendirmek için kullanılan kritik bir metriktir. Bu metrik, özellikle ASV ve CM sistemleri gibi uygulamalarda önemli bir role sahiptir. EER, sistemin iki temel hata türü olan *yanlış kabul oranı (FAR)* ve *yanlış reddetme oranı (FRR)* değerlerinin eşitlendiği noktada hesaplanır. Bu durum, sistemin hata oranlarını dengelediği ve belirli bir karar eşiği altında iki tür hatayı aynı oranda yaptığı anlamına gelmektedir.

İkili sınıflandırma sistemleri, genellikle bir karar skoru (örneğin benzerlik skoru veya doğrulama skoru) üretir ve karar verme işlemi bu skorun belirli bir eşik değer (threshold) ile karşılaştırılarak gerçekleştirilir. Bu eşik değeri, sistemin kabul veya reddetme performansını doğrudan etkilemektedir. Eğer eşik değer çok küçük seçilirse, sahte kullanıcıların kabul edilme olasılığı (FAR) artarken eğer eşik değer çok yüksekse, gerçek kullanıcıların reddedilme olasılığı (FRR) artmaktadır. EER, bu iki hatanın eşit olduğu eşik değerde hesaplanarak sistemin genel performansı hakkında bilgi veren bir metriktir.

Teorik olarak, FAR ve FRR değerleri her eşik seviyesinde hesaplanır ve bu değerler şekil 3.8'de görüldüğü gibi grafiksel olarak birbiriyle ilişkilendirilir. ROC eğrisi (Receiver Operating Characteristic), bu sürecin görsel bir ifadesidir. ROC eğrisi üzerinde FAR ve FRR'nin eşitlendiği nokta, EER'yi temsil etmektedir. Matematiksel olarak, bu noktada karar eşiği, sistemin hem güvenlik (FAR değerini minimum yapma) hem de erişilebilirlik (FRR değerini minimum yapma) hedeflerini dengelediği kritik bir değerdir.



Şekil 3.8: Eşit hata oranı grafiği.

EER değerinin düşük olması, sistemin her iki hata türünde de başarılı olduğunu ve daha tutarlı bir performans sergilediğini göstermektedir. Örneğin, bir konuşmacı doğrulama sisteminde düşük EER değeri, sistemin hem yetkisiz konuşmacıları reddetmede hem de yetkili konuşmacıları kabul etmede iyi performans gösterdiğini ifade etmektedir. Bu nedenle EER, sistemler arası karşılaştırma için tek bir sayısal değer sağlayarak kolaylıkla kullanılabilen bir başarım ölçütüdür.

Teorik olarak, ideal bir sistemde EER değeri sıfıra yaklaşır; bu durumda sistem ne yanlış kabul ne de yanlış reddetme hatası yapar. Ancak pratikte, sistemin eğitildiği veri setinin boyutu, kalite çeşitliliği ve kullanılan saldırı türleri gibi faktörler EER'nin düşürülmesini zorlaştırabilir. Bu nedenle, özellikle sahtekarlık farkındalığına sahip sistemlerde (SASV), EER kritik bir değerlendirme metriği olarak kullanılmaktadır.

3.6.2 a-DCF

a-DCF [86], detection cost function (DCF)'nin bir uzantısı olarak, yanıltma saldırılarına karşı dayanıklı konuşmacı doğrulama sistemlerinin performansını değerlendirmek için kullanılır. Bu metrik, geleneksel DCF'nin yalnızca iki hata türü (yanlış reddetme ve yanlış kabul) üzerinden hesaplama yapmasına kıyasla, üçüncü bir sınıf olan sahte (spoofing) sınıfını da değerlendirme sürecine dahil eder. Bu yapı, özellikle SASV (Spoofing Aware Speaker Verification) sistemlerinde etkin bir değerlendirme aracı sağlar. Teorik olarak a-DCF, bir sistemin üç temel hata türünü dikkate alarak hesaplanır:

1. *Target Miss (Yanlış Reddetme):* Gerçek kullanıcıların yanlışlıkla reddedilmesi.
2. *Non-Target False Acceptance (Yetkisiz Kabul):* Yetkisiz bir konuşmacının kabul edilmesi.
3. *Spoof False Acceptance (Sahte Kabul):* Sahte bir konuşmanın (spoofed speech) sistem tarafından kabul edilmesi.

a-DCF, bu üç hata türünü birleştirerek şu formülle tanımlanır [86]:

$$\begin{aligned} \text{a} - \text{DCF}(\tau_{\text{SASV}}) = \; & C_{\text{miss}}^{\text{tar}} . \pi_{\text{tar}} . P_{\text{miss}}^{\text{tar}}(\tau_{\text{SASV}}) \\ & + C_{\text{fa}}^{\text{non}} . \pi_{\text{non}} . P_{\text{fa}}^{\text{non}}(\tau_{\text{SASV}}) \\ & + C_{\text{fa}}^{\text{spf}} . \pi_{\text{spf}} . P_{\text{fa}}^{\text{spf}}(\tau_{\text{SASV}}) \end{aligned} \tag{3.15}$$

Denklem 3.15'te gösterilen:

- $C_{\text{miss}}^{\text{tar}}$: Yanlış reddetmenin maliyeti (miss cost),
- $C_{\text{fa}}^{\text{non}}$: Yetkisiz konuşmacının kabul edilme maliyeti,
- $C_{\text{fa}}^{\text{spf}}$: Sahte bir konuşmanın kabul edilme maliyeti,
- $\pi_{\text{tar}}, \pi_{\text{non}}, \pi_{\text{spf}}$: Hedef, yetkisiz ve sahte konuşmacıların olasılıkları,
- $P_{\text{miss}}^{\text{tar}}$: hedef kullanıcıya ait konuşmaların yanlışlıkla reddedilme oranını,

- $P_{\text{fa}}^{\text{non}}$: hedef olmayan bir kullanıcıya ait konuşmaların yanlışlıkla kabul edilme oranı,
- $P_{\text{fa}}^{\text{spf}}$: sahte (spoof) saldırılarına ait konuşmaların yanlışlıkla kabul edilme oranı,
- τ_{SASV}: Karar eşiği.

a-DCF, sistemin yalnızca karar eşiklerini değil, aynı zamanda maliyet parametrelerini ve olasılıkları da dikkate alarak hesaplama yapar. Böylece hem güvenlik hem de kullanıcı erişilebilirliği açısından sistemin performansı optimize edilir.

a-DCF metodu temelde aşağıdaki avantajları sağlamaktadır:

1. *Üçlü Hata Modeli***:** Sahte konuşmalar için ayrı bir hata oranını dahil ederek daha kapsamlı bir değerlendirme sağlar.
2. *Eşik Bağımsızlığı***:** Geleneksel t-DCF (tandem DCF)'nin aksine, sistem mimarisinden bağımsız olarak tek bir eşik üzerinden performansı değerlendirir.
3. *Güçlü Performans Değerlendirme***:** ASV ve CM sistemlerinin birlikte çalıştığı SASV sistemlerinde geniş kapsamlı performans analizi sağlar.

3.7 Önerilen Yöntemler

Bu kısımda, tez çalışması kapsamında gerçekleştirilen çalışmalar, çalışmaların başlıkları ile birlikte verilecek olan her bir çalışma alt başlık olarak sunulacaktır.

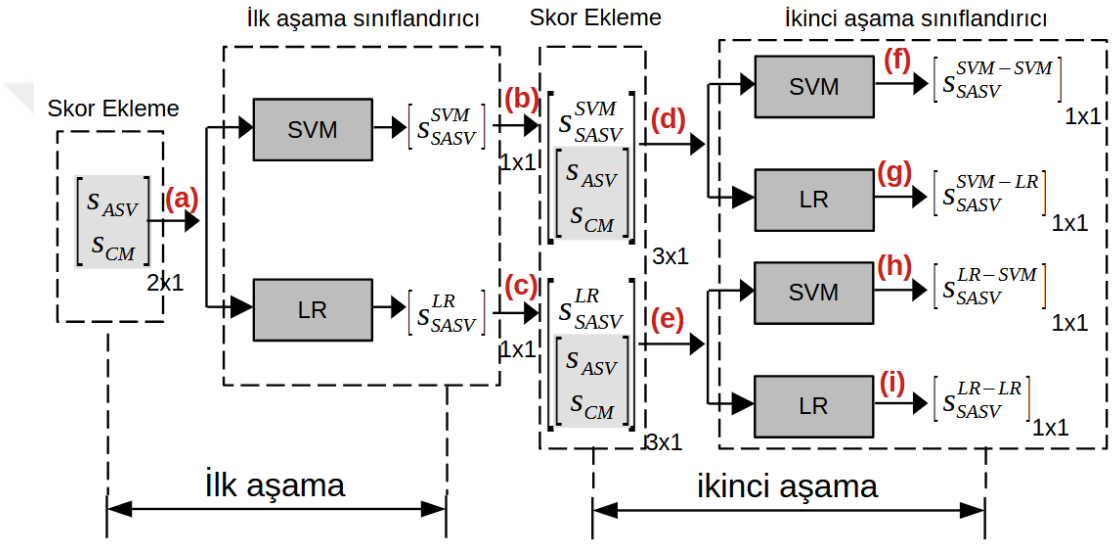
3.7.1 Saldırıdan haberdar konuşmacı doğrulamada çok aşamalı skor birleştirme yönteminin araştırılması

Bu çalışmada, sahtekarlık farkındalığına sahip konuşmacı doğrulama sistemlerinin performansını artırmak için modüler birçok aşamalı skor birleştirme yöntemi önerilmiştir. ASV ve CM alt sistemlerinin bağımsız olarak işlenip birden fazla aşamada birleştirilmesiyle, sistemin hem yanıltma saldırı tespitinde hem de konuşmacı doğrulamada daha hassas ve dayanıklı sonuçlar üretebilmesi hedeflenmiştir. Literatürde yaygın olarak kullanılan skor birleştirme yöntemlerinin uyum sağlama ve bilgi kaybı gibi sınırlamalarını aşmayı amaçlayan bu yöntem, ardışık aşamalarda veri entegrasyonunu optimize ederek daha sağlam bir performans sunar. Önerilen sistem, destek vektör makineleri veya lojistik regresyon gibi sınıflayıcılarla alt sistemlerden

gelen bilgileri entegre eder ve bu entegrasyonun SASV görevindeki potansiyelini ortaya koyar.

3.7.1.1 Kendinden artırımlı çok aşamalı skor birleştirme

Kendinden artırımlı çok aşamalı birleştirme yöntemi, şekil 3.9'da gösterildiği gibi, sırasıyla ECAPA-TDNN ve AASIST tarafından üretilen ASV ve CM skorlarını entegre eder. Daha sonra, SASV skorlarını türetmek için bir sınıflandırıcı eğitilir ve bu aşamadan sonra orijinal ASV ve CM skorlarıyla birlikte birleştirilmiş SASV skorlarını kullanarak başka bir sınıflandırıcı eğitilir.



Şekil 3.9: Kendinde artırımlı çok aşamalı skor birleştirme [122].

Bu birleştirme yöntemi, denklem 3.16'da gösterildiği gibi iki farklı alt birleştirme sistemini tanıtmaktadır: *iki skorlu önerilen çok aşamalı birleştirme* ve *üç skorlu önerilen çok aşamalı birleştirme*. İki skorlu yöntem, iki başlangıç skorunun (ASV ve CM) birleştirilmesini içerir. Burada, ECAPA-TDNN'den elde edilen ASV ve AASIST'ten elde edilen CM skorları, sonrasında türetilen SASV skorlarıyla birleştirilir. Üç skorlu yöntem ise iki skorlu yöntemin yapısını aynen korur, ancak üç başlangıç skorunu (ECAPA-TDNN'den bir ASV ve AASIST ile RawGAT-ST'den iki CM) kullanır. Önerilen iki skorlu yöntem şu şekilde formüle edilmiştir:

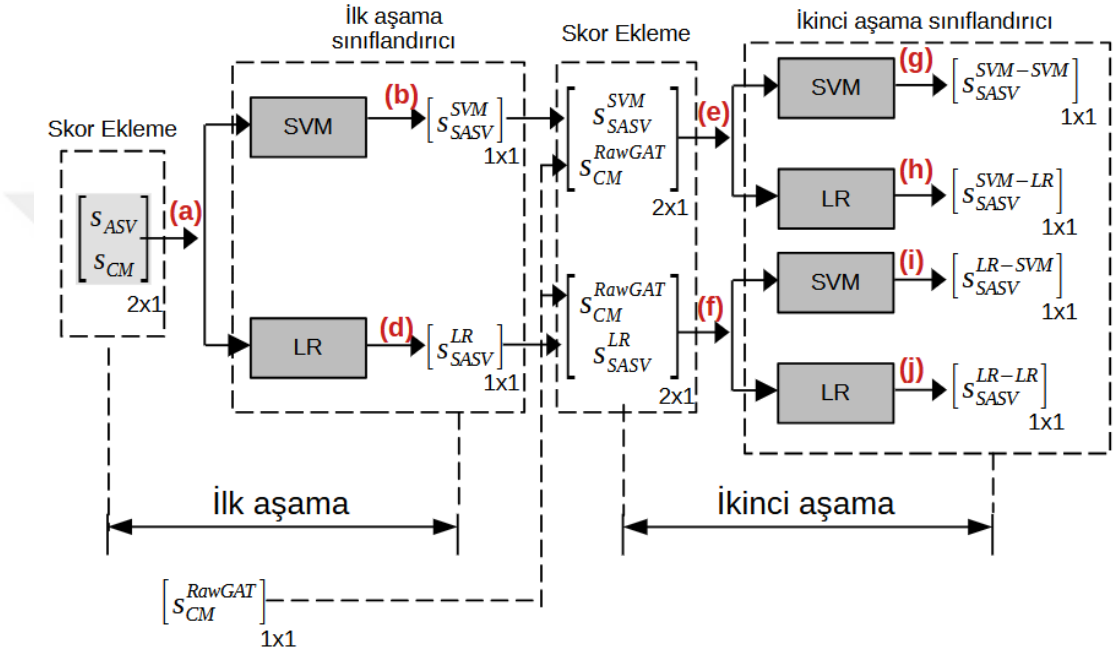
$$s = C_2^j\left(\left[C_1^i([s_{ASV}^{ECAPA}, s_{CM}^{AASIST}, s_{CM}^{RawGAT}]), s_{ASV}^{ECAPA}, s_{CM}^{AASIST}, s_{CM}^{RawGAT}\right]\right) \quad (3.16)$$
$$i, j \in \{SVM, LR\}$$

Burada, s_{ASV}^{ECAPA} ve s_{CM}^{AASIST}, sırasıyla başlangıç ASV ve CM skorlarını temsil etmektedir. C_1^i ve C_2^j, birinci ve ikinci aşama sınıflandırıcılarını ifade eder; burada i ve

j, SVM veya LR sınıflandırıcılarını belirtir. Önerilen yöntemde, birinci ve ikinci aşamalarda kullanılan birleştirme modellerine (SVM veya LR) bağlı olarak, Şekil 3.4'te gösterildiği gibi dört farklı iki aşamalı birleştirme stratejisi ele alınmıştır.

3.7.1.2 Harici artırımlı çok aşamalı skor birleştirme

Şekil 3.10'da gösterilen harici artırımlı çok aşamalı birleştirme yöntemi, önerilen çok aşamalı birleştirme yöntemiyle benzer bir konsepte sahiptir.



Şekil 3.10: Harici artırımlı çok aşamalı skor birleştirme [122].

İki başlangıç skorunu kullanan kendinden artırımlı yöntemden farklı olarak, denklem 3.17'de gösterilen harici artırımlı yaklaşım üç farklı skoru kullanır. Hem üç skorlu kendinden artırımlı yöntem hem de harici artırımlı yöntem üç skoru kullanma fikrini benimsemektedir; ancak birincisi, üç başlangıç skoru (ECAPA-TDNN'den bir ASV ve AASIST ile RawGAT-ST'den iki CM) kullanırken, ikincisi iki başlangıç skoruyla (ECAPA-TDNN'den ASV ve AASIST'ten CM) başlar ve sonraki aşamalarda ikinci CM (RawGAT-ST) skorunu sınıflayıcılar tarafından elde edilen SASV skorlarıyla birleştirir. Harici artırımlı çok aşamalı birleştirme yöntemi şu şekilde formüle edilebilir:

$$s = C_2^j\left(\left[C_1^i([s_{ASV}^{ECAPA}, s_{CM}^{AASIST}]), s_{CM}^{RawGAT}\right]\right) \quad i,j \in \{SVM, LR\} \tag{3.17}$$

Harici artırımlı çok aşamalı birleştirme yöntemi, önerilen çok aşamalı birleştirme yöntemine benzer şekilde, birinci ve ikinci aşamalarda seçilen sınıflandırıcılara bağlı olarak dört farklı iki aşamalı birleştirme varyasyonunu içermektedir.

3.7.2 Saldırıdan haberdar konuşmacı doğrulama için paralel derin öznitelik vektörü birleştirme yöntemi

Bu çalışmada ECAPA-TDNN, WavLM ve AASIST modellerinden elde edilen skorlar ve embedding vektörleri kullanılmıştır. Skor birleştirme yöntemleriyle a-DCF metrikleri hesaplanmış ve SASV performansı değerlendirilmiştir. Ayrıca, embedding birleştirme yöntemleriyle, farklı türdeki ASV ve CM embeddinglerini birleştirerek Baseline 2 modelini eğitmişlerdir. Ancak, farklı türdeki embeddinglerin birleştirilmesi, heterojen girdilere yol açtığı için tek bir DNN modeli bu verilerden etkili bir şekilde öğrenmekte zorlanabilir. Bu nedenle, bu çalışmada her biri farklı giriş verilerine odaklanan iki paralel DNN modeli önerilmiştir. Her model, kendi giriş özelliklerindeki özgün öznitelikleri öğrenerek SASV skorları üretmiş ve bu skorların ortalaması alınarak nihai SASV skoru elde edilmiştir. Bu yöntem, SASV sistemleri için daha hassas ve etkili sonuçlar sunmayı hedeflemiştir.

3.7.2.1 Skor birleştirme

Bir enrollment (kayıt) ve bir test konuşma çifti, s^{e} ve s^{t}, verildiğinde, SASV sistemi, s^{t}'yi $t_{SASV} \in \{0, 1\}$ olacak şekilde sınıflandırmayı amaçlar. Burada $t_{\mathrm{SASV}} = 1$, s^{e} ve s^{t}'nin aynı konuşmacıya ait olduğu ve s^{t}'nin gerçek (bonafide) olduğu durumu temsil ederken, $t_{\mathrm{SASV}} = 0$, s^{e} ve s^{t}'nin farklı konuşmacılardan geldiği ya da s^{t}'nin sahte (spoof) olduğu durumu ifade eder. Genel olarak, bir SASV sistemi oluşturmak için bağımsız olarak iki sınıflandırıcı geliştirilir: biri ASV görevi için, diğeri ise CM görevi için.

ASV sistemi, kayıt (s^{e}) ve test (s^{t}) konuşmalarından derin konuşmacı öznitelik vektörleri ($e_{\mathrm{ASV}}^{\mathrm{enr}}$ ve $e_{\mathrm{ASV}}^{\mathrm{tst}}$) çıkarır. Daha sonra, kayıt ve test öznitelik vektörleri arasındaki benzerlik derecesini temsil etmek için bu derin öznitelik vektörler kullanılarak denklem 3.18'de gösterilen kosinüs benzerliği hesaplanır. Ancak, kosinüs benzerliği değerleri [-1, 1] aralığında olduğundan, bu değerler bir sigmoid fonksiyonundan geçirilerek [0, 1] aralığına uyarlanır ve olasılıksal bir ASV skoru olarak işlenir:

$$P_{\text{ASV}} = P(t_{\text{ASV}} = 1 | e_{\text{ASV}}^{\text{enr}}, e_{\text{ASV}}^{\text{tst}}) = \sigma\left(\frac{(e_{\text{ASV}}^{\text{enr}})^{\text{T}} . (e_{\text{ASV}}^{\text{tst}})}{\|e_{\text{ASV}}^{\text{enr}}\| . \|e_{\text{ASV}}^{\text{tst}}\|}\right) \quad (3.18)$$

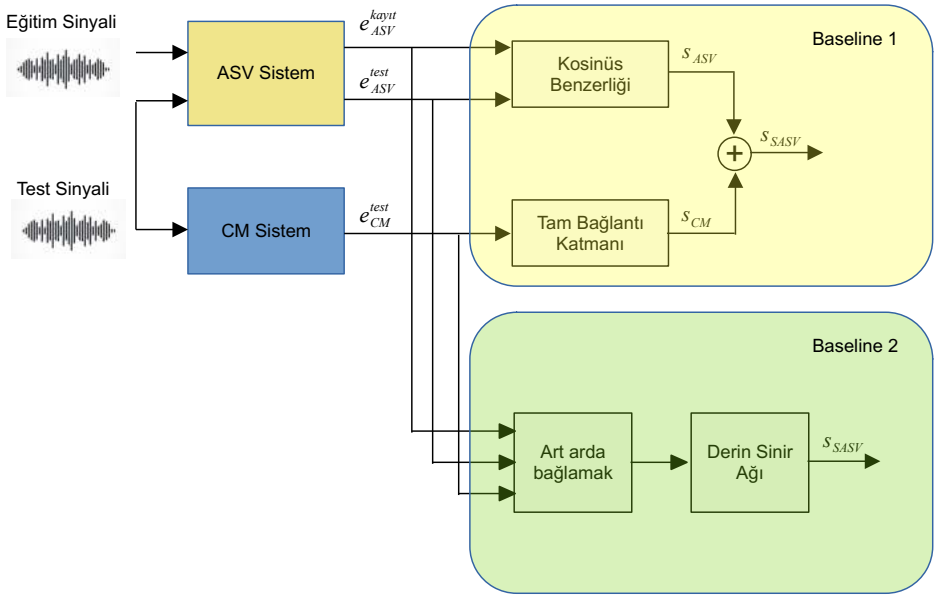
Burada σ, sigmoid fonksiyonudur ve $t_{\text{ASV}} \in \{0,1\}$ ile $t_{\text{ASV}} = 1$, kayıt ve test embeddinglerinin aynı konuşmacıya ait olduğunu, aksi takdirde $t_{\text{ASV}} = 0$'ı ifade eder. CM görevi için ise, CM sistemi bir CM derin öznitelik vektörü ($e_{\text{CM}}^{\text{tst}}$) çıkarır ve bu embedding, $t_{\text{CM}} \in \{0,1\}$ olacak şekilde sınıflandırılır. Burada $t_{\text{CM}} = 1$, test konuşmasının gerçek (bonafide) olduğunu, $t_{\text{CM}} = 0$ ise test konuşmasının sahte olduğunu gösterir. Bu sınıflandırma, bir softmax veya sigmoid aktivasyon fonksiyonunun çıktısı gibi olasılıksal bir şekilde CM skoru kullanılarak denklem 3.19'daki gibi gerçekleştirilir.

$$P_{\text{CM}} = P(t_{\text{CM}} = 1 | e_{\text{CM}}^{\text{tst}}) \quad (3.19)$$

SASV perspektifinden bakıldığında, $t_{\text{SASV}} = 1$ yalnızca $t_{\text{ASV}} = 1$ ve $t_{\text{CM}} = 1$ durumunda geçerlidir. Bu nedenle, toplam olasılık yasasına göre, ASV ve CM sistemlerinin çıktı olasılıkları, bir SASV sistemi oluşturmak için denklem 3.20'deki gibi birleştirilebilir.

$$P_{\text{SASV}} = P(t_{\text{SASV}} = 1 | e_{\text{ASV}}^{\text{enr}}, e_{\text{ASV}}^{\text{tst}}, e_{\text{CM}}^{\text{tst}}) = \frac{1}{2}(P_{\text{ASV}} + P_{\text{CM}}) \quad (3.20)$$

Bu yöntem, SASV2022 yarışması için önerilen Baseline 1 skor birleştirme yöntemine benzer şekilde tanımlanmıştır ve şekil 3.11'de özetlenmiştir.



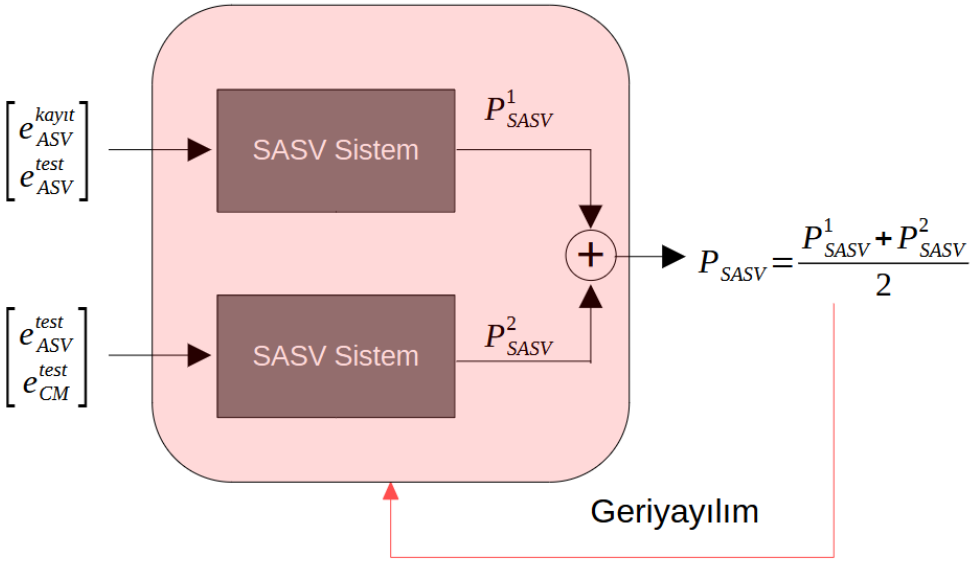
Şekil 3.11: SASV2022 yarışmasında verilen Baseline1 ve Baseline2 modelleri [59].

3.7.2.2 Derin öznitelik vektörü birleştirme

Temel embedding birleştirme yaklaşımında, ASV ($e_{\text{ASV}}^{\text{enr}}, e_{\text{ASV}}^{\text{tst}}$) ve CM ($e_{\text{CM}}^{\text{tst}}$) modellerinden elde edilen embeddingler birleştirilerek tek bir yüksek boyutlu özellik vektörü oluşturulur. Daha sonra, bu özellik vektörü, şekil 3.11'de gösterildiği gibi, birkaç tam bağlantılı katmandan (fully connected layers) oluşan basit bir DNN modeli kullanılarak $t_{\text{SASV}} \in \{0, 1\}$ olacak şekilde sınıflandırılır. Bu yapı, SASV2022 yarışmasında önerilen Baseline 2 sistemine benzer bir yapıdadır ve birleştirilen embeddingleri işleyerek P_{SASV} SASV skorunu üretir. Bu skorlar, sistemin performansını değerlendirmek için kullanılır ve farklı model derin öznitelik vektörlerinin birleşik gücünden yararlanılarak doğruluk ve dayanıklılık artırılır.

3.7.2.3 Paralel derin öznitelik vektörü birleştirme

Önceki bölümde açıklanan derin öznitelik vektörü birleştirme yöntemi, SASV görevi için makul bir yaklaşım olsa da tek bir DNN sisteminin, yığılmış ASV ve CM derin öznitelik vektörlerinden SASV problemi için en temsil edici özellikleri öğrenmekte zorlanabileceği belirtilmiştir. Bunun temel sebebi, bu yöntemin mimari sınırlamalarından kaynaklanmaktadır. Yapılan bazı ön deneyler, derin öznitelik vektörü birleştirme için kullanılan ASV ve CM derin öznitelik vektörlerinin farklı kombinasyonlarının farklı SASV performansları elde ettiğini göstermektedir. Bu durum, aynı DNN modelinin SASV sistemini eğitmek için kullanılmış olsa bile, giriş özelliklerine bağlı olarak farklı seviyelerde SASV bilgisi öğrenebildiğini ortaya koymaktadır. Bu nedenle, derin öznitelik vektörü birleştirme yaklaşımına dayalı, eşzamanlı çalışan ve birlikte optimize edilen iki paralel DNN'den oluşan bir SASV sistemi öneriyoruz.



Şekil 3.12: Paralel Derin Öznitelik Vektörü Birleştirme yöntemi [123].

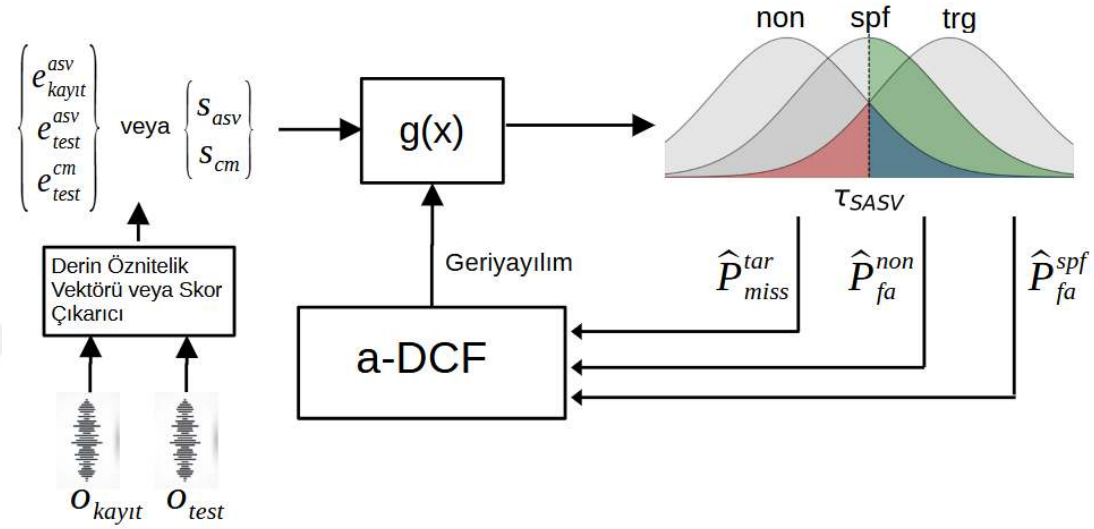
Şekil 3.12'de gösterilen model, SASV için tasarlanan önerilen paralel modeli göstermektedir. Bu modelin temel konsepti, mevcut ASV ve CM derin öznitelik vektörlerini kullanarak iki paralel çalışan özdeş sinir ağına beslemektir. Bu ağlar, aynı katman ve nöron sayısına sahip olmasına rağmen, farklı girişler alabilir ve bu da farklı giriş boyutlarına yol açabilir. Bu yapı, modelin çeşitli bilgi kaynaklarından faydalanmasını sağlarken, işleme tutarlılığını korur.

Bu paralel mimaride, birleştirilmiş derin öznitelik vektörler her bir SASV ağına beslenir. Her iki ağ da kendi girişlerini bağımsız olarak işler ve çıkışlarında SASV olasılıkları üretir. Bu ağların çıktıları sırasıyla P^1_{SASV} ve P^2_{SASV} olarak ifade edilir. Nihai SASV olasılığını elde etmek için bu bireysel olasılıkların ortalaması alınır. Bu ortalama işlemi, nihai SASV olasılığı olan P_{SASV}'nin her iki ağın çıktılarının dengeli bir şekilde değerlendirilmesini sağlar ve bu da SASV sisteminin dayanıklılığını artırma potansiyeline sahiptir.

3.7.3 Saldırıdan haberdar konuşmacı doğrulama için a-DCF kayıp fonksiyonu

Bir ASV ve CM sisteminin performansını değerlendirirken, yanlış reddetme (false rejection) ve yanlış kabul (false acceptance) oranları arasındaki denge dikkate alınmalıdır. Bu, kullanıcı deneyimi ve güvenlik arasındaki ödünleşmeyi ifade eder. Bu amaçla, tandem detection cost function (t-DCF) metriği, ASV ve CM kademelerinin Bayes risk çerçevesinde değerlendirilmesi için kullanılmıştır. Ancak, t-DCF iki farklı skor seti ve eşik gerektirdiğinden, diğer mimarilerin değerlendirilmesinde yetersiz

kalır. Bu sınırlamayı aşmak için, daha geniş bir mimari yelpazesinde uygulanabilen, mimariden bağımsız bir detection cost function (a-DCF) metriği önerilmiştir. a-DCF, yalnızca tek bir skor seti ve eşik gerektirerek t-DCF'den farklıdır ve ASVspoof 5 yarışmasının ana değerlendirme metriklerinden biri olarak kullanılmıştır.



Şekil 3.13: Önerilen a-DCF kayıp fonksiyonu modeli [78].

Şekil 3.13'te görseli sunulan bu çalışmada, ilk kez a-DCF metriği için doğrudan optimize edilen, yanıltma saldırılarına karşı dayanıklı bir ASV sistemi geliştirilmiştir. Ancak, a-DCF'nin sert hata sayımına dayalı olması, onu türevlenemez hale getirir ve dolayısıyla doğrudan gradyan inişi ile optimize edilemez. Fakat bu sert hata sayımları "yumuşatılarak" türevlenebilir hale getirilebilir. t-DCF'nin doğrudan optimizasyonu daha önce takviye öğrenme yöntemleriyle ele alınmış olsa da bu yaklaşımlar yalnızca tandem mimarilere uygulanabilir. Bu çalışmada ise, türevlenebilir bir a-DCF versiyonu model eğitimine entegre edilmiş ve ASV sistem performansı ile dayanıklılığı artırılmıştır. Ayrıca, a-DCF metriği ile sistem eşiği (τ_{SASV}) model parametreleriyle birlikte optimize edilmiştir.

3.7.3.1 DCF ve a-DCF

Detection Cost Function (DCF) [77], geleneksel ASV sistemlerinin performans değerlendirmesi için kullanılır. Bu metrik, yanlış alarmlar (false alarms) ve yanlış reddetmelerin (misses) sonuçlarını bir maliyet modeli üzerinden değerlendirir. DCF, denklem 3.21'deki gibi tanımlanmaktadır:

$$\mathrm{DCF}(\tau_{\mathrm{ASV}}) = C_{\mathrm{miss}}^{\mathrm{tar}} . \pi_{\mathrm{tar}} . P_{\mathrm{miss}}^{\mathrm{tar}}(\tau_{\mathrm{ASV}}) + C_{\mathrm{fa}}^{\mathrm{non}} . \pi_{\mathrm{non}} . P_{\mathrm{fa}}^{\mathrm{non}}(\tau_{\mathrm{ASV}}) \qquad (3.21)$$

Burada, $C_{\text{miss}}^{\text{tar}}$ ve $C_{\text{fa}}^{\text{non}}$ sırasıyla bir yanlış reddetme (miss) ve bir yanlış alarmın (FA) maliyetlerini temsil eder; $P_{\text{miss}}^{\text{tar}}$ ve $P_{\text{fa}}^{\text{non}}$ ise yanlış reddetme ve yanlış alarm oranlarını ifade eder. π_{tar} ve π_{non}= 1 - π_{tar}, hedef ve hedef dışı öncelik olasılıklarını temsil ederken, τ_{ASV} algılama eşik değeridir. Yukarıdaki denklemde verilen DCF, maliyet parametrelerinin belirli uygulama ihtiyaçlarına göre ayarlanmasına olanak tanır.

Mimariye bağımlı olmayan tespit maliyet fonksiyonu (a-DCF) [86], yanıltma saldırılarına karşı dayanıklı ASV sistemlerini değerlendirmek için kullanılır. Bu metrik, denklem 3.22'de gösterildiği gibi DCF'yi genişleterek üçüncü bir sınıfı (yanıltma saldırısı) da değerlendirme sürecine dahil eder:

$$\begin{aligned} \text{a} - \text{DCF}(\tau_{\text{SASV}}) = \; & C_{\text{miss}}^{\text{tar}} . \pi_{\text{tar}} . P_{\text{miss}}^{\text{tar}}(\tau_{\text{SASV}}) \\ & + C_{\text{fa}}^{\text{non}} . \pi_{\text{non}} . P_{\text{fa}}^{\text{non}}(\tau_{\text{SASV}}) \\ & + C_{\text{fa}}^{\text{spf}} . \pi_{\text{spf}} . P_{\text{fa}}^{\text{spf}}(\tau_{\text{SASV}}) \end{aligned} \tag{3.22}$$

Burada, $P_{\text{fa}}^{\text{spf}}$, yanıltma saldırı için yanlış alarm oranını; π_{spf}, yanıltma saldırı öncelik olasılığını; $C_{\text{fa}}^{\text{spf}}$, yanıltma saldırı yanlış alarm maliyetini ve τ_{SASV}, eşik değerini ifade eder. Denklem 3.22'de yer alan üç hata oranları denklem 3.23'te gösterildiği şekilde hesaplanmaktadır:

$$\begin{aligned} P_{\text{miss}}^{\text{tar}}(\tau_{\text{SASV}}) &= \frac{1}{N_{\text{tar}}} \sum_{x \in \text{tar}} I(g(x) \leq \tau_{\text{SASV}}) \\ P_{\text{fa}}^{\text{non}}(\tau_{\text{SASV}}) &= \frac{1}{N_{\text{non}}} \sum_{x \in \text{non}} I(g(x) > \tau_{\text{SASV}}) \\ P_{\text{fa}}^{\text{spf}}(\tau_{\text{SASV}}) &= \frac{1}{N_{\text{spf}}} \sum_{x \in \text{spf}} I(g(x) > \tau_{\text{SASV}}) \end{aligned} \tag{3.23}$$

Burada, tar, non ve spf sırasıyla hedef, hedef dışı ve sahte deneme kümelerini ifade eder ve bu kümelerdeki deneme sayıları N_* ile gösterilir.

3.7.3.2 Türevlenebilir a-DCF metriği

Denklem 3.23'te gösterilen indikatör fonksiyonu türevlenemezdir. Bu nedenle, a-DCF metriği, gradyan arama yöntemlerinde doğrudan bir kayıp fonksiyonu olarak kullanılamaz. Bu çalışmada, bu amaca uygun, türevlenebilir (yumuşak) bir a-DCF versiyonu önermekteyiz. Benzer yaklaşımlar, metne bağımlı konuşmacı doğrulama için [81] ve tandem mimariler için t-DCF'nin [56] optimize edilmesinde başarıyla

kullanılmıştır. Tandem sistemlere özgü gereksiz kısıtlamalar içeren t-DCF ile karşılaştırıldığında, a-DCF, keyfi SASV sistemlerine uygulanabilir.

a-DCF metriğindeki türevlenemeyen bileşenler, yanlış reddetme oranı ($P_{\text{miss}}^{\text{tar}}$) ve iki yanlış alarm oranıdır ($P_{\text{fa}}^{\text{non}}$, $P_{\text{fa}}^{\text{spf}}$). Sert hata sayımlarını "yumuşatmanın" yaygın bir yaklaşımı [81], [56], gösterge fonksiyonunu sürekli ve türevlenebilir bir fonksiyon ile değiştirmektir. Bu çalışmada, denklem 3.11'de gösterilen yaklaşımlar benimsenmiştir:

$$\hat{P}_{\text{miss}}^{\text{tar}}(\tau_{\text{SASV}}) = \frac{1}{N_{\text{tar}}} \sum_{x \in \text{tar}} \sigma\big(\tau_{\text{SASV}} - g(x)\big)$$

$$\hat{P}_{\text{fa}}^{\text{non}}(\tau_{\text{SASV}}) = \frac{1}{N_{\text{non}}} \sum_{x \in \text{non}} \sigma(g(x) - \tau_{\text{SASV}}) \tag{3.24}$$

$$\hat{P}_{\text{fa}}^{\text{spf}}(\tau_{\text{SASV}}) = \frac{1}{N_{\text{spf}}} \sum_{x \in \text{spf}} \sigma(g(x) - \tau_{\text{SASV}})$$

Burada $\sigma(z) = \frac{1}{1+e^{-z}}$ sigmoid fonksiyonudur. Temelde, hata oranları eşikten olan uzaklıklarına dayanarak yaklaşık olarak hesaplanır. Soft a-DCF kaybı, denklem 3.23'teki hata oranı terimlerinin, denklem 3.24'teki ifadelerle değiştirilmesi ve bu ifadelerin denklem 3.22'deki orijinal a-DCF formülüne uygulanmasıyla elde edilir.

3.7.3.3 Soft a-DCF optimizasyonu

Algorithm 1 Optimizasyon Algoritması

Girdi: Eğitim verisi $\mathcal{D}_{\text{trn}}$, Geliştirme verisi $\mathcal{D}_{\text{dev}}$, Mini-batch boyutu B, Epok sayısı N
Çıktı: En iyi eşik τ_{sasv}, Model Parametreleri Θ_*
Başlat: min_a−DCF $\leftarrow \infty$, $\tau \leftarrow 0.5$
for i $= 1$ to N **do**
    // DNN ağırlıklarını optimize et
    for $X \leftarrow get\_minibatch(\mathcal{D}_{\text{trn}}, B)$ **do**
        $\mathcal{J}(\Theta_i, \tau_i) \leftarrow \left(\mathcal{L}^{\text{soft}}_{\text{a-DCF}}(X; \Theta_i, \tau_i) + \mathcal{L}_{\text{BCE}}(X; \Theta_i)\right)/2$
        $\Theta_{i+1} \leftarrow update\_network\_weights(\Theta_i, \mathcal{J}(\Theta_i, \tau_i))$
    end for
    $\Theta \leftarrow \Theta_{i+1}$
    // grid search kullanarak optimal eşiği bul
    $\tau_{i+1} \leftarrow \text{argmin}_{\tau}\{\mathcal{L}^{\text{soft}}_{\text{a-DCF}}(\Theta, \tau)\}$ in $\mathcal{D}_{\text{trn}}$
    // geliştirici verisiyle en iyi modeli takip et
    if $\mathcal{L}^{\text{soft}}_{\text{a-DCF}}(\tau_{i+1}) <$ min_a−DCF in $\mathcal{D}_{\text{dev}}$ **then**
        min_a−DCF $\leftarrow \mathcal{L}^{\text{soft}}_{\text{a-DCF}}(\tau_{i+1})$
        $\tau_{\text{sasv}} \leftarrow \tau_{i+1}$
        $\Theta_* \leftarrow \Theta$
    end if
end for
return τ_{sasv}, Θ_*

Şekil 3.14: Soft a-DCF optimizasyon algoritması

Algoritma 1, yumuşatılmış a-DCF ($L^{\text{soft}}_{\text{a-DCF}}(X; \Theta_i, \tau_i)$) ve BCE ($L_{\text{BCE}}(X; \Theta_i)$) kayıplarının bir kombinasyonu aracılığıyla genel bir modelin optimizasyon döngüsünü sunar. İlk kayıp fonksiyonu olan ikili çapraz entropi (BCE) kayıp fonksiyonu denklem 3.25'teki gibi tanımlanmaktadır:

$$L_{\text{BCE}} = -\frac{1}{N}\sum_{j=1}^{N}\left[y_j \log\left(g(x_j)\right) + (1 - y_j)\log\left(1 - g(x_j)\right)\right] \tag{3.25}$$

Burada N, eğitim örneklerinin sayısını, $g(x_j)$ ise model çıktısını (skor) ifade eder. Algoritma, minimum a-DCF değerini sonsuz ve sabit bir eşik (τ) ile başlatır. Daha sonra, her epok için, yumuşatılmış a-DCF ve BCE kayıplarından birleşik kayıp $J(\Theta_i, \tau_i)$'yi hesaplayarak model ağırlıklarını optimize eder ve sinir ağının

ağırlıklarını günceller. Bu birleşik kayıp, şekil 3.14'te görülen Algoritma 1'in 7. satırında tanımlanmıştır.

Algoritma daha sonra, eğitim verilerindeki a-DCF kaybına göre optimal eşiği sağlayan τ değerini bulmak için bir grid arama işlemi gerçekleştirir (Algoritma 1'in 12. satırı). Eşik değeri, a-DCF kayıp fonksiyonunun denklem 3.24'te gösterilen hata oranlarında bağımsız değişken olarak hizmet eder. Model eğitimi sırasında hedef fonksiyon olarak a-DCF kaybı (veya BCE kaybı ile kombinasyonu) kullanıldığında, bu kayıp değerinin minimize edilmesi hedeflenir ve bu, eşik değerine bağlıdır. Bu ilişki, her epokta a-DCF kaybını minimize eden optimal eşik değerinin bulunmasını gerektirir ve bu işlem grid arama ile gerçekleştirilir.

Algoritma, geliştirme verisi üzerinde en iyi model parametrelerini ve eşik değerini izler; böylece şimdiye kadar bulunan en düşük a-DCF'yi ve buna karşılık gelen parametreleri günceller. Sonuç olarak, en iyi eşik değeri τ_{SASV}'yi döndürür.

3.7.3.4 Skor seviyesinde birleştirmede soft a-DCF

[79] çalışmasında, denklem 3.26'da gösterildiği gibi skor seviyesinde birleştirme yönteminin geliştirilmiş bir versiyonu olan lineer olmayan bir skor birleştirme yöntemi önerilmiştir.

$$s_{\text{SASV}} = -\log\left[(1-\rho)e^{-llr_{\text{non.bf}}^{\text{tar.bf}}(s)} + \rho e^{-llr_{\text{spf}}^{\text{tar.bf}}(s)}\right] \tag{3.26}$$

Denklem 3.26'da gösterilen eşitlikte ASV ve CM skorlarının log olabilirlik oranı (log likelihood ratio – LLR) lineer olmayan bir birleştirme yöntemi uygulanarak birleştirilmiştir. ASV ve CM skorların, logistic regresyon yöntemi uygulanarak kalibre edilmesi ve daha sonra lineer olmayan skor birleştirme yönteminin uygulanması ile standart skor birleştirme yöntemine göre daha başarılı sonuçlar elde edildiği görülmüştür. Bu sonuçlara referans alınarak, bu tez çalışmasında embedding birleştirme yöntemi için uygulanan soft a-DCF kayıp fonksiyonu, skor birleştirme için de uygulanmıştır. Embedding birleştirmede model çıkışları SASV skorları olarak kullanılırken, bu çalışmada kalibre edilmiş ASV ve CM skorlarının lineer olmayan birleştirilmesi ile elde edilen skorlar, SASV skorları olarak kullanılmıştır. Model optimizasyonu ise soft a-DCF kayıp fonksiyonu ve bu SASV skorları kullanılarak gerçekleştirilmiştir. Son durumda bu tez çalışmasında kullanılan kalibre edilmiş

skorların lineer olmayan birleştirilmesi işlemi denklem 3.27'deki gibi gerçekleştirilerek SASV skorları elde edilmiştir.

$$s_{\text{SASV}} = -\log\left[(1-\rho)e^{-(a*s_{\text{ASV}}+b)} + \rho e^{-(c*s_{\text{CM}}+d)}\right] \tag{3.27}$$

Burada $a, b, c,$ ve d değerleri, kalibrasyon katsayıları olup model içerisinde eğitilebilir parametreler olarak kullanılmaktadır. ρ değeri ise, sabit bir değer olup $\frac{\pi_{\text{non}}}{\pi_{\text{non}}+\pi_{\text{spf}}}$ olarak belirlenmiştir.

4. SONUÇLAR

Bu bölümde, tez çalışmasında gerçekleştirilen çalışmalara ilişkin sonuçlar önerilen yöntemler kısmındaki hiyerarşiye uygun olarak çalışma bazında verilecektir.

4.1 Saldırıdan haberdar konuşmacı doğrulamada çok aşamalı skor birleştirme yöntemi sonuçları

Önerilen çok aşamalı birleştirme sistemi, ASVspoof 2019 veri setinin Logical Access (LA) alt kümesi kullanılarak değerlendirilmiştir. Bu birleştirme yöntemi, polinom kernel kullanılan bir SVM sınıflandırıcı ve 10 katlı çapraz doğrulama (Cross Validation – CV) yöntemiyle çalışan lojistik regresyon (LR) sınıflandırıcılarını içermektedir.

SASV2022 yarışması kapsamında sunulan başlangıç modelleri, CM için ASVspoof 2019 veri setinin LA eğitim alt kümesinde eğitilmiş AASIST modeline ve ASV için VoxCeleb2 veri setinde eğitilmiş ECAPA-TDNN modeline dayanmaktadır. Bu çalışmada, ECAPA-TDNN ile AASIST ve RawGAT-STmodellerinden elde edilen ASV ve CM skorları kullanılmıştır.

Değerlendirme sürecinde iki farklı metrik kullanılmıştır. SASV sistemlerinin performansını karşılaştırmak amacıyla, hedef (pozitif) sınıfı, negatif sınıfın farklı alt kümelerinden ayırma yeteneğini ölçmek için üç farklı eşit hata oranı (EER) hesaplanmıştır: SV-EER, SPF-EER ve SASV-EER metrikleri sırasıyla hedef dışı, yanıltma saldırıları ve hedef dışı ile yanıltma saldırılarının bir karışımını negatif sınıf olarak kullanmıştır. Buna ek olarak, ASV araştırmalarında yaygın olarak kullanılan klasik DCF'yi üç sınıflı değerlendirme için genişleten yeni bir mimariden bağımsız algılama maliyet fonksiyonu (a-DCF) metriği de kullanılmıştır. a-DCF, üç farklı EER raporlamak yerine, tek bir değer sunar: minimum normalize edilmiş algılama maliyeti. Bu çalışmada, hedef, hedef dışı ve yanıltma saldırısı öncelikleri sırasıyla 0,9, 0,05 ve 0,05; hedef kaçırma, hedef dışı yanlış alarm ve yanıltma saldırı yanlış alarm maliyetleri

ise sırasıyla 1, 10 ve 20 olacak şekilde ayarlanmış açık kaynak bir uygulama kullanılarak hesaplamalar gerçekleştirilmiştir.

4.1.1 Tüm sistem sonuçları

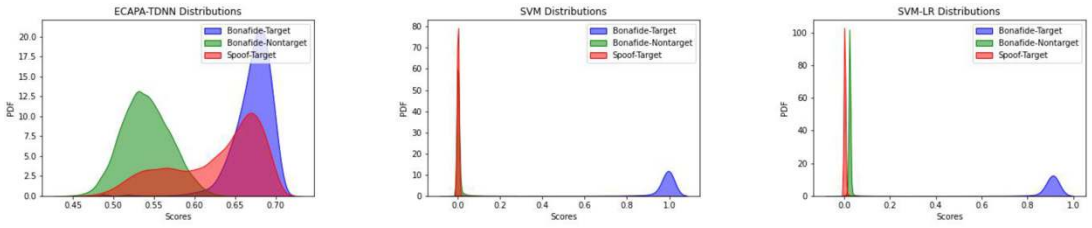
Önerilen çok aşamalı skor seviyesi birleştirme çerçevesine ait deneysel sonuçlar çizelge 3'te özetlenmiştir. "Aşama" sütunu her skorun kaynak aşamasını, "Birleştirme" sütunu ise uygulanan birleştirme stratejilerini belirtir. "Sistem" sütunu, kullanılan sistemi ifade eder. Çok aşamalı bölümde, dört harften oluşan kombinasyonlar, ilgili sistem skorlarını elde etmek için şekil 3.9 ve şekil 3.10'da gösterilen yolları temsil eder. "Sınıflandırıcı" sütunu, şekil 3.9 ve şekil 3.10'da skor üretimi için kullanılan sınıflandırıcıları belirtir. Ayrıca, "Kullanılan skorlar" sütunu hangi modellerden elde edilen skor özniteliklerinin kullanıldığını ifade eder; burada E, A ve R harfleri sırasıyla ECAPA-TDNN, AASIST ve RawGAT-ST modellerinden elde edilen skorları temsil eder. Baseline sistem, basit bir skor toplama birleştirme tekniği kullanarak değerlendirme veri setinde %1,71 SASV-EER elde etmiş ve orta düzeyde bir performans sergilemiştir. Tek aşamalı birleştirme yaklaşımlarına geçildiğinde, üç skoru kullanan (a)-LR ve (a)-SVM yöntemleri ile performans daha da iyileşmiştir ve sırasıyla %1,55 ve %1,38 SASV-EER, 0,033 ve 0,029 a-DCF değerleri elde edilmiştir.

Çizelge 3: Çok aşamalı skor seviyesi birleştirme sonuçları.

Aşama	Birleştirme	Sistem	Sınıflandırıcı	Kullanılan Skorlar	SASV-EER		a-DCF	
					Gel.	Değ.	Gel.	Değ.
-	Skor toplama	Baseline B1	-	E, A	1.01	1.71	N/A	N/A
Birinci	Tek aşama	(a)-LR	C_1^{LR}	E, A	1.15	1.61	0.030	0.036
Birinci	Tek aşama	(a)-SVM	C_1^{SVM}	E, A	0.94	1.60	0.027	0.038
Birinci	Tek aşama	(a)-LR	C_1^{LR}	E, A, R	1.08	1.55	0.029	0.033
Birinci	Tek aşama	(a)-SVM	C_1^{SVM}	E, A, R	0.94	1.38	0.027	0.029
İkinci	Ken. Art. çok aşama	(a)-(c)-(e)-(i)	$C_2^{LR}(C_1^{LR})$	E, A	1.17	1.58	0.031	0.035
İkinci	Ken. Art.çok aşama	(a)-(c)-(e)-(h)	$C_2^{SVM}(C_1^{LR})$	E, A	1.82	2.36	0.034	0.034
İkinci	Ken. Art.çok aşama	(a)-(b)-(d)-(g)	$C_2^{LR}(C_1^{SVM})$	E, A	1.00	1.60	0.028	0.037
İkinci	Ken. Art.çok aşama	(a)-(b)-(d)-(f)	$C_2^{SVM}(C_1^{SVM})$	E, A	2.49	2.53	0.032	0.032
İkinci	Ken. Art.çok aşama	(a)-(c)-(e)-(i)	$C_2^{LR}(C_1^{LR})$	E, A, R	1.08	1.43	0.029	0.031
İkinci	Ken. Art.çok aşama	(a)-(c)-(e)-(h)	$C_2^{SVM}(C_1^{LR})$	E, A, R	2.09	2.09	0.030	0.030
İkinci	Ken. Art.çok aşama	(a)-(b)-(d)-(g)	$C_2^{LR}(C_1^{SVM})$	E, A, R	0.95	1.30	0.027	0.028
İkinci	Ken. Art.çok aşama	(a)-(b)-(d)-(f)	$C_2^{SVM}(C_1^{SVM})$	E, A, R	2.16	2.03	0.029	0.028
İkinci	Har. Art. çok aşama	(a)-(d)-(f)-(j)	$C_2^{LR}(C_1^{LR})$	E, A, R	1.15	1.53	0.030	0.031
İkinci	Har. Art.çok aşama	(a)-(d)-(f)-(i)	$C_2^{SVM}(C_1^{LR})$	E, A, R	1.82	1.96	0.031	0.029
İkinci	Har. Art.çok aşama	(a)-(b)-(e)-(h)	$C_2^{LR}(C_1^{SVM})$	E, A, R	0.94	1.53	0.027	0.035

İkinci	Har. Art.çok aşama	(a)-(b)-(e)-(g)	$C_2^{SVM}(C_1^{SVM})$	E, A, R	2.36	2.33	0.030	0.029

Kendinde artırımlı çok aşamalı birleştirme yöntemlerine geçiş yapıldığında, farklı birleştirme yolları ve stratejileri arasında performans metriklerinde belirgin farklılıklar gözlemlenmiştir. Özellikle ikinci aşamada, çok aşamalı birleştirme yapılandırmaları arasında, üç skoru kullanan (a)-(b)-(d)-(g) yolu dikkat çekmiştir. Bu yapılandırma, değerlendirme veri setinde %1,30 SASV-EER ve 0,028 a-DCF ile etkileyici bir performans sergilemiştir. Bu sonuç, sıralı birleştirme aşamalarının SASV performansını iyileştirmede ve optimize etmede önemli bir potansiyele sahip olduğunu vurgulamaktadır. Benzer bir gözlem, şekil 4.1'de dört farklı özellikle elde edilen skor dağılımlarından da yapılabilir. (a)-(b)-(d)-(g) yolunun SASV görevinde güçlü ayrıştırıcı yeteneklere sahip olduğu açıkça görülmektedir. Örneğin, *ECAPA-TDNN* (Şekil 4.1 (a)) hedef ve hedef dışı sınıflarını etkili bir şekilde ayırt edebilirken, yanıltma saldırı sınıfının dağılımı geniş bir aralık göstermekte ve hem hedef hem de hedef dışı sınıflarla örtüşmektedir. Buna karşılık, en iyi sistemimizin skor dağılımları (Şekil 4.1 (c)) üç sınıfın diğer durumlara kıyasla daha iyi ayrıştığını göstermektedir. Şekil 4.1'de verilen dağılımlarda 'Bonafide-Target' etiketi hedef (Target) sınıfını belirtmekte olup, 'Bonafide-Nontarget' hedef olmayan (Nontarget), 'Spoof-Target' ise sahte (Spoof) sınıfını ifade etmektedir.



(a) ECAPA-TDNN (b) Single-stage SVM (c) Multi-stage SVM-LR

Şekil 4.1: (a) ECAPA-TDNN, (b) tek aşamalı SVM ve (c) çok aşamalı SVM-LR sınıflandırıcıların skor dağılımları [122].

Harici artırımlı çok aşamalı birleştirme bölümünde, hem (a)-(d)-(f)-(j) hem de (a)-(b)-(e)-(h) yolları, önerilen iki skorlu çok aşamalı birleştirmeye kıyasla iyileştirmeler sağlamıştır ve %1,53 SASV-EER ile sırasıyla 0,031 ve 0,035 a-DCF değerlerine ulaşmıştır. Genel olarak, sonuçlar, çok aşamalı birleştirme tekniklerinin kullanımının avantajlarını vurgulamakta ve sahtekarlık farkındalığına sahip konuşmacı doğrulama sistemlerinin dayanıklılığını ve güvenilirliğini artırmadaki etkinlikleri hakkında değerli bilgiler sunmaktadır.

4.1.2 Saldırı tipi bazında sonuçlar

Çizelge 4'te sunulan her bir saldırı türüne ait bireysel sonuçlar, sistem performansının farklı saldırı senaryolarında detaylı bir şekilde değerlendirilmesini sağlamaktadır. Bu çizelgede "Sistem" sütunu, kullanılan sistemi belirtir. A07-A19 arasındaki sütunlar saldırı türlerini, "pooled" sütunu ise tüm saldırı türlerinin birleştirildiğini gösterir. Listelenen sistemler, performanslarına göre çizelge 3'ten seçilmiştir. ECAPA-TDNN sistemi, neredeyse tüm saldırı türlerinde yüksek hata oranları göstermektedir. Buna karşılık, AASIST sistemi, A09 saldırısında oldukça iyi bir performans sergilemekte ve A10 ile A16 saldırılarında en iyi sonuçlara ulaşmaktadır, ancak diğer saldırı türlerinde zorlanmaktadır. Baseline-B1 ve (a)-SVM sistemleri, genel olarak düşük hata oranlarını korumaktadır, ancak belirli saldırılar için hafif artışlar gözlemlenmiştir.

Çizelge 4: Herbir saldırı için elde edilen EER değerleri.

Sistem	A07	A08	A09	A10	A11	A12	A13	A14	A15	A16	A17	A18	A19	Pooled
ECAPA	32.7	18.8	2.2	50.6	47.1	39.6	11.6	35.4	36.5	60.7	1.9	2.4	4.8	30.75
AASIST	0.80	0.44	**0.00**	**1.06**	0.31	0.91	0.10	0.14	0.65	**0.72**	1.52	3.40	0.62	1.13
B1	2.05	0.69	0.07	7.19	0.32	4.42	**0.07**	0.08	1.75	1.18	0.73	1.18	0.63	0.78
(a)-SVM	0.78	0.65	0.04	1.62	0.19	1.12	0.10	0.08	0.80	1.20	0.88	1.49	0.73	0.93
(a)-(c)-(e)-(i)	0.95	0.35	**0.00**	2.95	0.11	1.73	**0.07**	0.08	0.92	0.96	0.58	0.99	0.41	1.02
(a)-(b)-(d)-(g)	**0.26**	**0.34**	**0.00**	1.14	**0.07**	**0.45**	**0.07**	**0.06**	**0.31**	1.00	**0.51**	1.21	0.39	**0.59**
(a)-(d)-(f)-(j)	0.33	0.37	0.04	1.91	0.26	1.10	0.26	0.26	0.35	0.93	0.63	**0.93**	**0.37**	0.76

Çok aşamalı birleştirme yöntemleri, özellikle (a)-(c)-(e)-(i), (a)-(b)-(d)-(g) ve (a)-(d)-(f)-(j) yolları, rekabetçi performans sergilemiştir. Özellikle (a)-(b)-(d)-(g) yolu, dikkat çekici derecede düşük hata oranlarına ulaşmıştır. Bu bulgular, çok aşamalı birleştirme tekniklerinin, sistem dayanıklılığını farklı saldırı türleri karşısında artırmadaki etkinliğini vurgulamaktadır.

4.2 Saldırıdan haberdar konuşmacı doğrulama için paralel derin öznitelik vektörü birleştirme yöntemi

Deneylerde ASV derin öznitelik vektörleri çıkarmak için iki model kullanılmıştır: (i) VoxCeleb2 veri seti kullanılarak eğitilen ECAPA-TDNN modeli ve (ii) önceden eğitilmiş WavLM modeli. ECAPA-TDNN modeli, kayıt ve test konuşmalarından 192 boyutlu ASV derin öznitelik vektörleri çıkarmaktadır. Bu derin öznitelik vektörler, ya ASV skorunu hesaplamak için denklem 3.3'te, derin öznitelik vektörü birleştirme yönteminde veya önerilen paralel SASV sisteminde kullanılmaktadır. WavLM modeli

ise konuşmacı kimliğini koruyarak konuşma içeriğini modellemeye odaklanır. Bu modelin eğitimi, örtüşen konuşma parçalarının oluşturulmasını içeren konuşma karışım stratejisini içerir. Bu çalışmada, WavLM'nin önceden eğitilmiş modeli kullanılarak, her bir konuşma için 768 boyutlu özellik temsilleri çıkarılmıştır. Bu temsiller, konuşma başına ortalama havuzlama (mean pooling) yöntemiyle işlenerek derin öznitelik vektörü seviyesinde ASV görevine entegre edilmiştir.

CM sistemi için AASIST modeli kullanılmış ve bu model ASVspoof5 eğitim alt kümesi ile eğitilmiştir. AASIST, ham ses dalga formlarını işleyerek yüksek boyutlu spektral-zamansal özellik haritalarını öğrenmektedir. Daha sonra bu haritalardan, zaman ve frekans alanlarındaki düğümleri çıkarmaktadır. Bu düğümlerden elde edilen ortalama ve maksimum değerler birleştirilerek, 160 boyutlu CM derin öznitelik vektörleri oluşturulmuştur. Bu derin öznitelik vektörler hem temel model hem de önerilen birleştirme çerçevesinde SASV görevini gerçekleştirmek için kullanılmıştır.

SASV sistemi kapsamında, temel modelde (şekil 3.11), derin öznitelik vektörü birleştirme yapılmış ve bu derin öznitelik vektörler bir DNN yapısına giriş olarak verilmiştir. Bu yapı, bir giriş katmanı, üç gizli katman (256, 128 ve 64 düğüm) ve bir çıkış katmanından oluşmaktadır. Önerilen paralel yapıda (şekil 3.12), derin öznitelik vektörüler iki paralel DNN yapısına giriş olarak verilmiştir. Bu yapılar bağımsız olarak eğitilmiş ve eğitimden sonra her iki DNN'nin çıktılarından elde edilen SASV skorları ortalanmıştır. Bu ortalama skor, paralel DNN yapılarını optimize etmek için kullanılmıştır. Paralel modeller, giriş boyutları dışında neredeyse tamamen aynıdır ve gizli katmanların düğüm sayıları temel modelle aynıdır. Bu yapı, SASV sistemlerinin performansını artırmayı ve sahtekarlığa dayanıklılığı geliştirmeyi hedeflemektedir.

Yapılan deneylerde, önerilen çoklu aşamalı birleştirme sisteminin performansı ASVspoof 5 veriseti üzerindeki geliştirme ve ilerleme alt kümelerinde değerlendirildi. Çizelge 5'te, minimum a-DCF değerleri farklı sistem yapılandırmalarıyla özetlenmiştir. Tabloda, kullanılan yöntem (skor birleştirme veya derin öznitelik vektörü birleştirme), kullanılan ASV sistemleri, CM sistemleri, uygulanan optimizasyon yöntemi ve geliştirme ile ilerleme setlerindeki a-DCF sonuçları yer almaktadır. ECAPA-TDNN, WavLM ve AASIST modellerinin farklı birleştirme stratejileri ve optimizasyon teknikleri altındaki kombinasyonları karşılaştırılmaktadır. Her alt grup içinde en küçük a-DCF değerleri kalın yazılmış, globalde en küçük değerler ise altı çizili ve kalın yazılmıştır. Basit skor birleştirme ve derin öznitelik

vektörü birleştirme yöntemleri başlangıçta kullanılmış, skor birleştirme yönteminin minimum a-DCF metriğinde derin öznitelik vektörü birleştirme yönteminden daha iyi performans gösterdiği tespit edilmiştir. Örneğin, skor birleştirme yöntemi değerlendirme setinde 0,3767 a-DCF değeri elde ederken, derin öznitelik vektörü birleştirme yöntemi 0,3956 a-DCF değeri elde etmiştir

Çizelge 5: Paralel derin öznitelik vektörü yöntemi ile elde edilen sonuçlar [123].

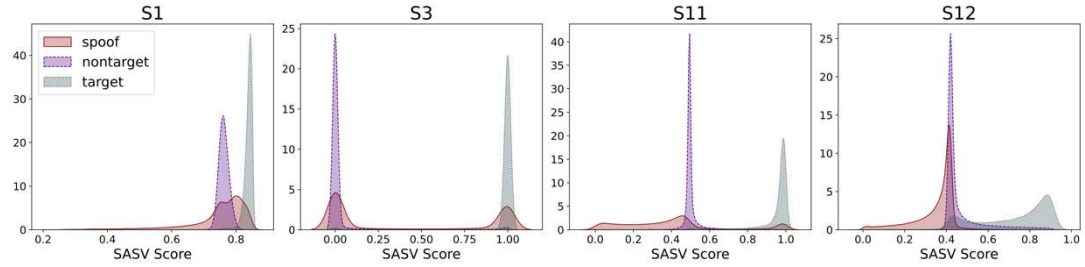
Sistem	Birleştirme Yöntemi	ASV Sistem	CM Sistem	Kayıp Fonksiyonu	min a-DCF	
					Gel.	Değ.
S1	Skor	ECAPA	AASIST	-	0.2268	0.3767
S2	Embedding	ECAPA	AASIST	a-DCF	0.2776	0.4046
S3	Embedding	ECAPA	AASIST	BCE	0.2830	0.3956
S4	Embedding	ECAPA	AASIST	a-DCF+BCE	**0.2754**	**0.3937**
S5	Embedding	WavLM	AASIST	a-DCF	**0.2677**	0.4084
S6	Embedding	WavLM	AASIST	BCE	0.2951	0.3927
S7	Embedding	WavLM	AASIST	a-DCF+BCE	0.2801	**0.3831**
S8	Embedding	ECAPA, WavLM	AASIST	a-DCF	0.2316	**0.3790**
S9	Embedding	ECAPA, WavLM	AASIST	BCE	0.2204	0.4364
S10	Embedding	ECAPA, WavLM	AASIST	a-DCF+BCE	**0.2088**	0.3949
S11	Paralel	ECAPA	AASIST	a-DCF+BCE	**0.1692**	**0.2492**
S12	Paralel	WavLM	AASIST	a-DCF+BCE	0.1820	0.3253
S13	Paralel	ECAPA, WavLM	AASIST	a-DCF+BCE	0.2482	0.3582
S14	Skor	ECAPA, WavLM	AASIST	-	**0.1250**	**0.2129**

Derin öznitelik vektörü birleştirme ağı eğitimi sırasında farklı kayıp fonksiyonlarının etkisi incelenmiştir. Geleneksel BCE kaybı, yalnızca a-DCF kaybı veya BCE ve a-DCF kayıplarının birleşimi ile değiştirilmiştir. Sonuçlar, BCE ve a-DCF kayıplarının birleşiminin (0,3937 a-DCF ile) tek başına kullanılan BCE (0,3956) ve a-DCF (0.4046) kayıplarından daha iyi performans sağladığını göstermiştir.

ASV derin öznitelik vektörüleri çıkarmak için ECAPA-TDNN ve WavLM modelleri kullanılmıştır. WavLM modeli, konuşmacıya özgü bilgileri koruyarak, ECAPA-TDNN'e kıyasla daha düşük a-DCF değerleri sağlamıştır. Örneğin, ECAPA-TDNN derin öznitelik vektörleriyle 0,3956 a-DCF elde edilirken, WavLM derin öznitelik vektörleriyle bu değer 0,3927'ye düşmüştür. Ayrıca, her iki modelin derin öznitelik vektörleri birleştirildiğinde, a-DCF değerleri önemli ölçüde iyileşmiştir. Bu birleşim,

%22 oranında göreceli bir iyileşme ile 0,2204 a-DCF elde etmiştir, bu da her iki modelin tamamlayıcı bilgiler sağladığını göstermektedir.

Son olarak, önerilen paralel model farklı ASV derin öznitelik vektörü çıkarıcıları ile test edilmiştir. Paralel model, BCE ve a-DCF kayıplarının birleşimini kullanarak, temel sistemlere kıyasla daha iyi performans göstermiştir. Özellikle, ECAPA-TDNN derin öznitelik vektörü çıkarıcısını kullanan iki katmanlı paralel model (128 ve 64 düğüm), geliştirme ve ilerleme setlerinde en iyi sonuçları sağlamıştır. Ayrıca, en iyi performans gösteren iki paralel modelin skorlarının birleştirilmesiyle %14 oranında bir performans artışı elde edilmiştir. Şekil 4.2'de, önerilen paralel modelin, hedef (target), hedef dışı (nontarget) ve sahte (spoof) sınıflarının skorlarını temel yöntemlere kıyasla daha iyi ayırdığı görülmüştür.



Şekil 4.2: Çizelge 5'te sonuçları verilen S1, S3, S11 ve S12 sistemlerinin skor dağılımları.

Geliştirme ve ilerleme alt kümelerinde gerçekleştirilen kapsamlı deneyler temelinde, kapalı koşul başvurusu için *ECAPA-TDNN* ve *AASIST* modellerinden elde edilen derin öznitelik vektörlerle önerilen paralel model (*S11 sistemi*) kullanılarak değerlendirme alt kümesindeki SASV skorları hesaplanmıştır. Açık koşul başvurusu için ise, *S11* ve *S12* paralel modellerinden elde edilen skorların (*S14 sistemi*) birleştirilmesi yöntemi uygulanmıştır. Değerlendirme setine ilişkin a-DCF değerleri çizelge 6'da özetlenmiştir.

Çizelge 5: S11 ve S14 sistemlerinin değerlendirme setindeki sonuçları.

Sistem	Birleştirme Yöntemi	min a-DCF (Değ. Set)
S11	Paralel	0.5130
S14	Skor (S11 + S12)	0.4581

ECAPA-TDNN ve AASIST derin öznitelik vektörlerini kullanan paralel modelle yapılan başvurumuz (S11), 0,5130 a-DCF skoru elde etmiştir. Buna karşılık, skor seviyesi birleştirme yöntemi daha iyi bir performans göstermiş ve 0,4581 a-DCF değeri elde edilmiştir. Bu eğilim, tüm veri alt kümelerinde tutarlı bir şekilde

gözlemlenmiş ve sahtekarlık farkındalığına sahip konuşmacı doğrulama için sağlam ve etkili bir yöntem geliştirdiğimiz sonucunu desteklemiştir.

4.3 Saldırıdan haberdar konuşmacı doğrulama için a-DCF kayıp fonksiyonu

Önerilen optimizasyon şemasını göstermek için, SASV2022 yarışması kapsamında [14]'te tanımlanan ve bir başlangıç noktası olarak kullanılan DNN tabanlı derin öznitelik vektörü birleştirme stratejisi 'Baseline 2'yi ele aldık. Daha sonra, birçok diğer çalışma benzer bir yaklaşımı izleyerek SASV gerçekleştirmiştir [42], [80]. İki konuşmacı derin öznitelik vektörü ($e_{\mathrm{ASV}}^{\mathrm{enr}}, e_{\mathrm{ASV}}^{\mathrm{tst}}$) ECAPA-TDNN [4] ile ve sahtekarlık derin öznitelik vektörü ($e_{\mathrm{CM}}^{\mathrm{tst}}$) AASIST [39] ile çıkarılmıştır. Üç derin öznitelik vektörü birleştirilerek, 256, 128 ve 64 nörondan oluşan üç tam bağlı gizli katmana sahip bir DNN modeline verilmiştir. Bu model, Leaky ReLU aktivasyon fonksiyonlarını kullanmaktadır. Baseline modelindeki iki nöronlu softmax çıkışı yerine, modelimizin son katmanı, yalnızca bir çıkış skoru gerektiren a-DCF optimizasyonu için bir sigmoid aktivasyon fonksiyonu ile tek bir nörona sahiptir.

Deneylerde, aynı mimarilere sahip ancak farklı şekilde optimize edilen dört arka uç modeli değerlendirilmiştir:

- *S1*: SASV2022 yarışma düzenleyicileri tarafından sağlanan Baseline 2 sistemi (CE ile optimize edilmiştir),
- *S2*: Yumuşak a-DCF kullanılarak, τ_{SASV} = 0,5 ile optimize edilmiştir,
- *S3*: Yumuşak a-DCF ve BCE'nin kombinasyonu ile, τ_{SASV} = 0,5 ile optimize edilmiştir,
- *S4*: Yumuşak a-DCF ve BCE'nin kombinasyonu ile ve Algoritma 1'de detaylandırıldığı gibi eşik optimizasyonu dahil edilerek optimize edilmiştir.

Bu derin öznitelik vektörü birleştirme yöntemlerine ek olarak, eğitilebilir bir doğrusal olmayan skor birleştirme yönteminin [79] a-DCF optimizasyonu önerilmiştir (detaylar aşağıda verilmiştir).

Yumuşak a-DCF hesaplamasında kullanılan maliyet değerleri $C_{\mathrm{miss}}^{\mathrm{tar}}$, $C_{\mathrm{fa}}^{\mathrm{non}}$ ve $C_{\mathrm{fa}}^{\mathrm{spf}}$ sırasıyla 1, 10 ve 20 olarak ayarlanmıştır. Öncelik değerleri π_{tar}, π_{non} ve π_{spf} ise sırasıyla 0,9, 0,05 ve 0,05 olarak belirlenmiştir. Sahtekarlık yanlış alarmlarının ($C_{\mathrm{fa}}^{\mathrm{spf}}$ = 20) yüksek göreceli maliyet değeri, özellikle bankacılık gibi güvenlik açısından

hassas alanlarda sahtekarlık dayanıklılığının önemine dayanmaktadır. Bu yaklaşım, sahte girişimlerin yanlış kabul edilmesini cezalandırarak sistemin bütünlüğünü ve güvenliğini sağlama amacını yansıtmaktadır. Ayrıca, yüksek hedef öncelik ($\pi_{\text{tar}} = 0{,}9$), çoğu etkileşimin meşru olduğu varsayımı ile uyumludur. Performans değerlendirmesi için Denklem (2)'de tanımlanan a-DCF metriği kullanılmıştır. Ayrıca, seçili durumlarda eşit hata oranı (EER) raporlanmıştır. SV-EER, sistemin hedef ve hedef dışı örnekleri ayırt etme yeteneğini değerlendirmek için, SPF-EER ise meşru hedefler ile sahtekarlık saldırıları arasındaki farkı ayırt etme yeteneğini değerlendirmek için kullanılmıştır.

Deneyler sırasında, öncelikle çapraz entropi kaybı (CE) kullanılarak Baseline (S1) modeli eğitilmiştir ve minimum a-DCF değerleri çizelge 7'de sunulmuştur. Sonuçlar, en iyi performansın 1024 batch boyutunda elde edildiğini göstermiştir ve bu değer sonraki deneyler için sabitlenmiştir.

Çizelge 6: Baseline 2 sisteminin geliştirme setinde batch boyutuna göre sonuçları.

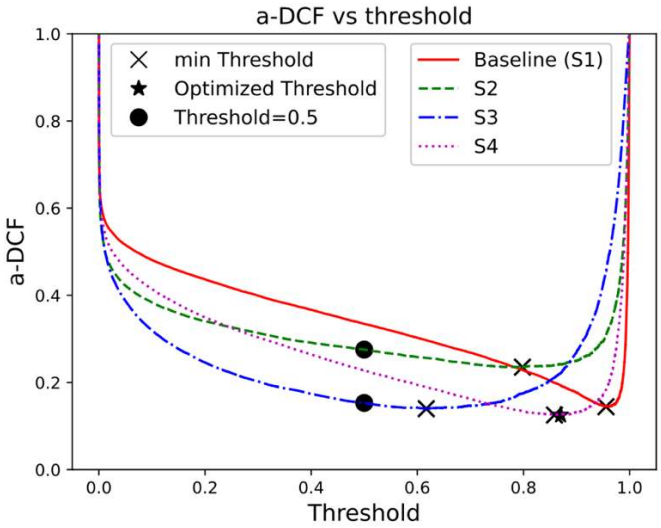
Batch Boyutu	24	48	96	192	384	768	1024	2048
a-DCF	0.1437	0.1308	0.1268	0.1346	0.1297	0.1251	0.1234	0.1243

Daha sonra, çizelge 8'de gösterildiği gibi S2 ve S3 modelleri, yumuşak a-DCF kaybı kullanılarak eşik değeriyle optimize edilmiştir. S2 modeli, geliştirme setinde %9,76 SV-EER ve %0,14 SPF-EER ile minimum 0,1355 a-DCF değerine ulaşmıştır. S3 modeli ise yumuşak a-DCF ve BCE kayıplarını birleştirerek bu değerleri daha da iyileştirmiş, geliştirme setinde 0,1182 a-DCF ve %8,29 SV-EER elde etmiştir. Eşik optimizasyonunun da dahil edildiği S4 modeli, parametrelerin ve eşiklerin ortak optimizasyonu sayesinde, geliştirme setinde 0,1109 a-DCF, %7,75 SV-EER ve %0,08 SPF-EER ile en iyi performansı göstermiştir.

Çizelge 7: S1 (Baseline), S2, S3 ve S4'ün geliştirme ve değerlendirme setindeki sonuçları.

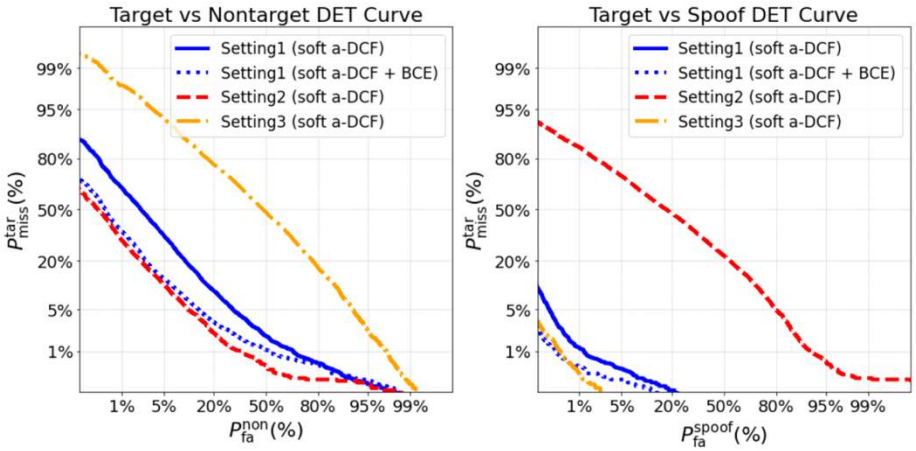
Sistem	a-DCF		SV-EER		SPF-EER	
	Gel.	Değ.	Gel.	Değ.	Gel.	Değ.
S1	0.1234	0.1445	8.36	9.86	0.07	0.63
S2	0.1355	0.2352	9.76	17.28	0.14	1.02
S3	0.1182	0.1398	8.29	9.59	0.10	0.63
S4	0.1109	0.1254	7.75	8.44	0.08	0.61

Şekil 4.3, S1, S2, S3 ve S4 sistemleri için farklı eşik değerleri boyunca a-DCF eğrilerini sunmaktadır. S2 ve S3 sistemleri için, hem sabit $\tau_{\text{SASV}} = 0{,}5$ eşik değerinde hem de minimum noktadaki a-DCF değerleri belirtilmiştir. Eşik optimizasyonunu içeren S4 sisteminin optimize edilmiş a-DCF değeri, grafikte bir yıldız ile işaretlenmiştir. Grafikten de görüldüğü üzere, S4 sistemi en düşük a-DCF değerini elde ederek diğer sistemlere kıyasla daha iyi performans göstermiştir.



Şekil 4.3: Farklı eşik değerlerinde Baseline (S1), S2, S3 ve S4 sistemlerinin a-DCF değerleri [78].

Yumuşak a-DCF parametrelerinin etkisini incelemek için üç farklı parametre ayarı değerlendirilmiştir. İlk ayar, yalnızca yumuşak a-DCF kaybını içerirken, ikinci ayar a-DCF ve BCE kayıplarının birleştirilmesini içermektedir. Diğer iki ayar (setting2 ve setting3), öncelik değerlerinde değişiklikler yaparak farklı senaryoları simüle etmiştir. Setting2'de maliyet değerleri $C_{\text{miss}}^{\text{tar}} = C_{\text{fa}}^{\text{non}} = C_{\text{fa}}^{\text{spf}} = 1$ ve öncelik değerleri $\pi_{\text{tar}} = \pi_{\text{non}} = 0{,}5$ ve $\pi_{\text{spf}} = 0$ şeklindedir. Setting3'te ise maliyet değerleri $C_{\text{miss}}^{\text{tar}} = C_{\text{fa}}^{\text{non}} = C_{\text{fa}}^{\text{spf}} = 1$ ve öncelik değerleri $\pi_{\text{tar}} = \pi_{\text{spf}} = 0{,}5$ ve $\pi_{\text{non}} = 0$ şeklindedir. Şekil 4.4'teki DET eğrileri, seçilen a-DCF parametrelerinin sistem performansı üzerindeki önemli etkisini açıkça göstermiştir. Bu eğriler üç farklı a-DCF parametre ayarı için verilmiştir. SV-EER değerleri, Ayar1 (düz mavi) için %13,97, Ayar1 (soft a-DCF + BCE) (noktalı mavi) için %8,44, Ayar2 için %7,52 ve Ayar3 için %49,13'tür. Örneğin sahte olmayan hedeflerin olmadığı bir durum (setting2), hedef ve hedef dışı denemeler arasında en iyi performansı sağlamıştır. Ancak, bu eğilimler hedef ve sahte denemeler için tersine dönmüştür.



Şekil 4.4: Üç farklı a-DCF ayarı için, Hedef-Hedef Olmayan ve Hedef-Sahtekarlık DET eğrileri [78].

Skor birleştirme (SF) için yumuşak a-DCF optimizasyonu da test edilmiştir. Çalışmada, doğrusal kalibre edilmiş skor birleştirme (L-Cal-SF) ve doğrusal olmayan kalibre edilmiş skor birleştirme (NL-Cal-SF) yöntemlerinin performansı karşılaştırılmıştır. Önerilen yöntem (aDCF-NL-Cal-SF), NL-Cal-SF'ye kıyasla minimum a-DCF değerini 0,0508'den 0,0289'a düşürmüştür. Ayrıca, skor birleştirme yönteminin, daha az parametre kullanarak (4 parametre ile 180736 parametreye karşı) derin öznitelik vektörü birleştirme yöntemine kıyasla daha iyi SASV performansı sağladığı gözlemlenmiştir. Bu sonuçlar, önerilen birleştirme tekniklerinin sahtekarlık farkındalığına sahip konuşmacı doğrulama sistemlerinde etkili bir şekilde kullanılabileceğini ortaya koymuştur.

Çizelge 9: Farklı SASV sistemlerinin önerilen skor seviyesinde birleştirme yöntemine göre karşılaştırılması.

SASV Sistemi	# Eğitilebilir Parametre	min a-DCF
Derin Öznitelik Sev. Birleştirme	180736	0.1234
L-SF	0	0.5311
L-Cal-SF	4	0.0648
NL-Cal-SF	4	0.0508
aDCF-NL-Cal-SF	4	0.0289

5. TARTIŞMA VE ÖNERİLER

Bu çalışmada, sahtekarlık farkındalığına sahip konuşmacı doğrulama (SASV) sistemlerinin geliştirilmesine odaklanılmış ve mevcut literatürdeki en ileri yöntemlerin bir sentezi gerçekleştirilmiştir. Çalışmanın temel amacı, konuşmacı doğrulama sistemlerinin güvenilirliğini artırmak ve sahtecilik saldırılarına karşı dayanıklılığını geliştirmektir. Gerek ECAPA-TDNN, WavLM ve AASIST gibi güçlü modellerin birleştirilmesi gerekse çok aşamalı skor seviyeli birleştirme ve paralel derin sinir ağı (DNN) mimarileri gibi yenilikçi yaklaşımlar, çalışmanın çerçevesini oluşturmuştur.

Yapılan deneyler, SASV sistemlerinde skor seviyesi birleştirmenin, derin öznitelik vektörü seviyesindeki birleştirme yöntemlerine kıyasla daha yüksek bir performans sunduğunu göstermiştir. Özellikle, çok aşamalı birleştirme stratejileri ile optimize edilen sistemlerin, minimum a-DCF ve SASV-EER değerlerinde önemli düşüşler sağladığı gözlemlenmiştir. Bu durum, farklı aşamalarda bilgi entegrasyonunun, sistemin genel performansını artırmada kritik bir rol oynadığını göstermektedir.

Bununla birlikte, WavLM ve ECAPA-TDNN modellerinin birleştirilmesiyle elde edilen sonuçlar, her iki modelin tamamlayıcı bilgi taşıdığını ve birlikte kullanıldığında sahtekarlık tespitinde belirgin bir iyileşme sağladığını ortaya koymuştur. Ancak, bu modellerin farklı veri kümeleri üzerinde eğitilmiş olması ve özelliklerinin homojen olmaması, sistemin genelleştirme kapasitesini sınırlayabilmektedir. Bu nedenle, daha iyi bir genelleme için, modellerin ortak bir özellik uzayı oluşturacak şekilde optimize edilmesi gerektiği düşünülmektedir.

Paralel DNN mimarileriyle yapılan deneylerde, her bir ağın farklı veri türlerini işlemek için optimize edilmesi, sahtekarlık farkındalığını artırmış ve sistem performansını iyileştirmiştir. Bununla birlikte, bu yaklaşımın hesaplama maliyetini artırdığı ve gerçek zamanlı uygulamalarda kullanımını zorlaştırabileceği gözlemlenmiştir. Ayrıca, optimize edilen a-DCF metriğinin kullanımı, sahtekarlık tespitinin performansını artırsa da bu metriklerin bazı durumlarda fazla cezalandırıcı olabileceği ve dolayısıyla sistemin belirli sahtecilik senaryolarına aşırı hassas hale gelebileceği anlaşılmıştır.

Bu çalışmanın bulguları doğrultusunda, aşağıdaki öneriler sunulabilir:

1. *Veri Çeşitliliği ve Genelleme Kapasitesinin Artırılması*: Mevcut SASV sistemlerinin performansı, eğitim sırasında kullanılan veri kümelerinin çeşitliliği ile doğrudan ilişkilidir. Daha geniş ve farklı türde sahtecilik saldırılarını içeren veri kümeleri ile eğitilmiş sistemlerin genelleme kapasitesi artırılabilir.

2. *Model Birleştirme Tekniklerinin Geliştirilmesi*: WavLM ve ECAPA-TDNN gibi modellerin birleştirilmesiyle elde edilen sonuçlar umut verici olmakla birlikte, bu modellerin heterojen doğası göz önünde bulundurularak daha etkili birleştirme stratejileri araştırılmalıdır. Özellikle, farklı modellerin ortak bir özellik uzayında optimize edilmesi, birleştirme sırasında oluşabilecek bilgi kaybını minimize edebilir.

3. *Hesaplama Maliyetinin Azaltılması*: Paralel DNN mimarileri güçlü performans sunmasına rağmen, gerçek zamanlı uygulamalarda bu mimarilerin maliyeti bir sınırlama olabilir. Bu nedenle, daha hafif ve hızlı modellerin geliştirilmesi, SASV sistemlerinin yaygın kullanımını teşvik edecektir.

4. *Metriğin Yeniden Değerlendirilmesi*: a-DCF metriği, sahtekarlık farkındalığına sahip doğrulama sistemlerinde etkili bir performans değerlendirme yöntemi sunmaktadır. Ancak, metriklerin uygulandığı senaryolara aşırı duyarlı olmaması için daha esnek ve genel bir değerlendirme çerçevesi oluşturulmalıdır.

Bu tez çalışması, mevcut literatürdeki yenilikçi tekniklerin bütünleştirilmesi ve optimize edilmesi yoluyla SASV sistemlerinin performansını ve güvenilirliğini artırmaya yönelik önemli bir katkı sunmaktadır. Ancak, daha geniş veri çeşitliliği, daha hafif model yapıları ve gerçek dünya senaryolarına uygunluk gibi alanlarda yapılacak çalışmalar, bu sistemlerin daha ileri seviyeye taşınmasını sağlayacaktır.

KAYNAKLAR

[1] **Reynolds, D. A.** (1995). Speaker identification and verification using Gaussian mixture speaker models. In *Speech Communication*, 17(1), 91–108.

[2] **Dellwo, V., Huckvale, M., Ashby, M.** (2007). How is individuality expressed in voice? An introduction to speech production and description for speaker classification. In *Speaker Classification I: Fundamentals, Features, and Methods*, (pp. 1–20). Springer.

[3] **Peddinti, V., Povey, D., Khudanpur, S.** (2015). A time delay neural network architecture for efficient modeling of long temporal contexts. In *Proceedings of INTERSPEECH 2015*, (pp. 3214–3218).

[4] **Desplanques, B., Thienpondt, J., Demuynck, K.** (2020). ECAPA-TDNN: Emphasized channel attention, propagation, and aggregation in TDNN-based speaker verification. In *Proceedings of INTERSPEECH 2020*, (pp. 3560–3564).

[6] **Chung, J. S., Huh, J., Mun, S., Lee, M., Heo, H. S., Choe, S., Ham, C., Jung, S., Lee, B.-J., Han, I.** (2020). In defence of metric learning for speaker recognition. In *Proceedings of INTERSPEECH 2020*, (pp. 2977–2981).

[7] **Li, L., Nai, R., Wang, D.** (2022). Real additive margin softmax for speaker verification. In *Proceedings of ICASSP 2022*, (pp. 7527–7531).

[8] **Reynolds, D. A.** (2002). An overview of automatic speaker recognition technology. In *Proceedings of ICASSP 2002*, (pp. 4072–4075).

[9] **Ge, W.** (2024). Spoofing-robust Automatic Speaker Verification: Architecture, Explainability and Joint Optimisation. In *Doctoral dissertation*, Sorbonne Université.

[10] **Sahidullah, M., Delgado, H., Todisco, M., Nautsch, A., Wang, X., Kinnunen, T., Evans, N., Yamagishi, J., Lee, K.-A.** (2023). Introduction to voice presentation attack detection and recent advances. In *Handbook of Biometric Anti-Spoofing: Presentation Attack Detection and Vulnerability Assessment*, (pp. 339–385).

[11] **Kinnunen, T., Sahidullah, M., Delgado, H., Todisco, M., Evans, N., Yamagishi, J., Lee, K.-A.** (2017). The ASVspoof 2017 Challenge: Assessing the limits of replay spoofing attack detection. In *Proceedings of INTERSPEECH 2017*, (pp. 2–6).

[12] **Frank, J., Schönherr, L.** (2021). WaveFake: A data set to facilitate audio deepfake detection. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

[13] **Todisco, M., Wang, X., Vestman, V., Sahidullah, M., Delgado, H., Nautsch, A., Yamagishi, J., Evans, N., Kinnunen, T. H., Lee, K.-A.** (2019). ASVspoof 2019: Future horizons in spoofed and fake audio detection. In *Proceedings of INTERSPEECH 2019*, (pp. 1008–1012).

[14] **Shim, H.-J., Tak, H., Liu, X., Heo, H.-S., Jung, J.-W., Chung, J. S., Chung, S.-W., Yu, H.-J., Lee, B.-J., Todisco, M., et al.** (2022). Baseline systems for the first spoofing-aware speaker verification challenge:

Score and embedding fusion. In *Proceedings of Odyssey 2022 - The Speaker and Language Recognition Workshop*.

[15] **Pruzansky, S., Mathews, M. V.** (1964). Talker-recognition procedure based on analysis of variance. *Journal of the Acoustical Society of America, 36*(11), 2041–2047. https://doi.org/10.1121/1.1919222.

[16] **Doddington, G. R.** (1971). A method for speaker verification. *Journal of the Acoustical Society of America, 49*(1A), 139–139. https://doi.org/10.1121/1.1912368.

[17] **Atal, B. S.** (1972). Text-independent speaker recognition. *Journal of the Acoustical Society of America, 52*(1A), 181–181. https://doi.org/10.1121/1.1912851.

[18] **Furui, S.** (1981). Cepstral analysis technique for automatic speaker verification. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 29*(2), 254–272. https://doi.org/10.1109/TASSP.1981.1163530.

[19] **Soong, F. K., Rosenberg, A., Rabiner, L., Juang, B.** (1987). A vector quantization approach to speaker recognition. *AT&T Technical Journal, 66*, 22–30. https://doi.org/10.1109/ICASS.1987.1163534.

[20] **Poritz, A.** (1982). Linear predictive hidden Markov models and the speech signal. In *ICASSP'82: IEEE International Conference on Acoustics, Speech, and Signal Processing* (pp. 1291–1294). IEEE. https://doi.org/10.1109/ICASSP.1982.1163547.

[21] **Matsui, T., Furui, S.** (1994). Similarity normalization method for speaker verification based on a posteriori probability. In *Proceedings of the ESCA Workshop on Automatic Speaker Recognition, Identification, and Verification* (pp. 59–62). Switzerland.

[22] **Higgins, A., Bahler, L., Porter, J.** (1991). Speaker verification using randomized phrase prompting. *Digital Signal Processing, 1*(2), 89–106. https://doi.org/10.1016/1051-2004(91)90019-B.

[23] **Reynolds, D. A., Quatieri, T. F., Dunn, R. B.** (2000). Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing, 10*(1–3), 19–41. https://doi.org/10.1006/dspr.2000.0360.

[24] **Kenny, P.** (2005). Joint factor analysis of speaker and session variability: Theory and algorithms. *CRIM Report CRIM-06/08-13*, 14(28–29), 2.

[27] **Snyder, D., Ghahremani, P., Povey, D., Garcia-Romero, D., Carmiel, Y., Khudanpur, S.** (2016). Deep neural network-based speaker embeddings for end-to-end speaker verification. In *Proceedings of the Spoken Language Technology Workshop*. https://doi.org/10.1109/SLT.2016.1234567.

[28] **Sahidullah, M., Delgado, H., Todisco, M., Yu, H., Kinnunen, T., Evans, N., Tan, Z.-H.** (2016). Integrated spoofing countermeasures and automatic speaker verification: An evaluation on ASVspoof 2015.

[29] **Kinnunen, T., Sahidullah, M., Delgado, H., Todisco, M., Evans, N., Yamagishi, J., Lee, K.-A.** (2017). The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection. In: INTERSPEECH. https://doi.org/10.21437/Interspeech.2017-1111.

[30] **Yang, J., Das, R. K.** (2019). Low frequency frame-wise normalization over constant-q transform for playback speech detection. *Digital Signal Processing*. https://doi.org/10.1016/j.dsp.2019.02.018.

[31] **Lai, C.-I., Abad, A., Richmond, K., Yamagishi, J., Dehak, N., King, S.** (2019). Attentive filtering networks for audio replay attack detection. In:

ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 6316–6320). IEEE.

[32] **Huang, L., Pun, C.-M.** (2020). Audio replay spoof attack detection by joint segment-based linear filter bank feature extraction and attention-enhanced DenseNet-BiLSTM network. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 28*, 1813–1825.

[33] **Tak, H., Jung, J.-W., Patino, J., Kamble, M., Todisco, M., Evans, N.** (2021). End-to-End Spectro-Temporal Graph Attention Networks for Speaker Verification Anti-Spoofing and Speech Deepfake Detection. *arXiv*. https://doi.org/10.48550/ARXIV.2107.12710.

[34] **Tapkir, P. A., Patil, H. A.** (2018). Significance of Teager energy operator phase for replay spoof detection. In: *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)* (pp. 1951–1956). IEEE.

[35] **Novoselov, S., Kozlov, A., Lavrentyeva, G., Simonchik, K., Shchemelinin, V.** (2016). STC anti-spoofing systems for the ASVspoof 2015 challenge. In: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5475–5479). IEEE.

[36] **Wu, Z., Das, R. K., Yang, J., Li, H.** (2020). Light convolutional neural network with feature genuinization for detection of synthetic speech attacks. *arXiv*. https://doi.org/10.48550/ARXIV.2009.09637.

[37] **Tapkir, P. A., Patil, A. T., Shah, N., Patil, H. A.** (2018). Novel spectral root cepstral features for replay spoof detection. In: *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)* (pp. 1945–1950). IEEE.

[38] **Wu, X., He, R., Sun, Z., Tan, T.** (2018). A light CNN for deep face representation with noisy labels. *IEEE Transactions on Information Forensics and Security, 13*(11), 2884–2896.

[39] **Jung, J.-W., Heo, H.-S., Tak, H., Shim, H.-J., Chung, J. S., Lee, B.-J., Yu, H.-J., Evans, N.** (2021). AASIST: Audio Anti-Spoofing using Integrated Spectro-Temporal Graph Attention Networks. *arXiv*. https://doi.org/10.48550/ARXIV.2110.01200.

[40] **Grinberg, P., Shikhov, V.** (2022). A Comparative Study of Fusion Methods for SASV Challenge 2022. *arXiv preprint arXiv:2203.16970*.

[41] **Ta, B. T., Nguyen, T. L., Dang, D. S., Le, D. L.** (2022, July). A multi-task conformer for spoofing aware speaker verification. In *2022 IEEE Ninth International Conference on Communications and Electronics (ICCE)* (pp. 306-310). IEEE.

[42] **Zhang, Y., Zhu, G., Duan, Z.** (2022). A probabilistic fusion framework for spoofing aware speaker verification. *arXiv preprint arXiv:2202.05253*.

[43] **Alenin, A., Torgashov, N., Okhotnikov, A., Makarov, R., Yakovlev, I.** (2022, September). A Subnetwork Approach for Spoofing Aware Speaker Verification. In *INTERSPEECH* (pp. 2888-2892).

[44] **Shim, H. J., Tak, H., Liu, X., Heo, H. S., Jung, J. W., Chung, J. S., Chung, S. W., Yu, H. J., Lee, B. J., Todisco, M., Nautsch, A., Evans, N., Yamagishi, J. (2022).** Baseline systems for the first spoofing-aware speaker verification challenge: Score and embedding fusion. *arXiv preprint arXiv:2204.09976.*

[45] **Ge, W., Tak, H., Todisco, M., Evans, N.** (2022). On the potential of jointly-optimised solutions to spoofing attack detection and automatic speaker verification. *IberSPEECH Conference*.

[46] **Kim, N., Kang, T., Choi, S., Shin, J., Oh, S., Kwak, I. (2022).** CAU KU team's system description for 2022 spoofing-aware speaker verification challenge. *In Proceedings of SASV Challenge 2022.*

[47] **Wu, H., Meng, L., Kang, J., Li, J., Li, X., Wu, X., … Meng, H.** (2022). Spoofing-aware speaker verification by multi-level fusion. *arXiv preprint arXiv:2203.15377*.

[48] **Liu, X., Sahidullah, M., Lee, K. A., Kinnunen, T.** (2024). Generalizing speaker verification for spoof awareness in the embedding space. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

[49**] Choi, J. H., Yang, J. Y., Jeoung, Y. R., Chang, J. H.** (2022, September). HYU Submission for the SASV Challenge 2022: Reforming Speaker Embeddings with Spoofing-Aware Conditioning. In *INTERSPEECH* (pp. 2873-2877).

[50] **Todisco, M, Delgado, H, Aik Lee, K, Sahidullah, M, Evans, N, Kinnunen, T Yamagishi, J.** (2018), Integrated Presentation Attack Detection and Automatic Speaker Verification: Common Features and Gaussian Back-end Fusion. in *Proc. Interspeech 2018.* Interspeech, International Speech Communication Association, Hyderabad, India, pp. 77-81, Interspeech 2018, Hyderabad, India, 2/09/18. https://doi.org/10.21437/Interspeech.2018-2289.

[51] **Shim, H. J., Jung, J. W., Kim, J. H., Yu, H. J.** (2020). Integrated replay spoofing-aware text-independent speaker verification. *Applied Sciences*, *10*(18), 6292.

[52] **Li, J., Sun, M., Zhang, X.** (2019). Multi-task learning of deep neural networks for joint automatic speaker verification and spoofing detection. *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 1517-1522.

[53] **Zhao, Y., Togneri, R., Sreeram, V.** (2022). *Multi-task Learning-Based Spoofing-Robust Automatic Speaker Verification System.* Circuits, Systems, and Signal Processing, 41(6), 3553–3568.

[54] **Zhang, P., Hu, P., Zhang, X.** (2022). Norm-constrained Score-level Ensemble for Spoofing Aware Speaker Verification. *Interspeech2022*.

[55] **Gomez-Alanis, A., Gonzalez-Lopez, J. A., Dubagunta, S. P., Peinado, A. M., Doss, M. M.** (2020). On joint optimization of automatic speaker verification and anti-spoofing in the embedding space. *IEEE Transactions on Information Forensics and Security*, *16*, 1579-1593.

[56] **Kanervisto, A., Hautamäki, V., Kinnunen, T., Yamagishi, J.** (2021). Optimizing tandem speaker verification and anti-spoofing systems. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 30*, 477-488.

[57] **Lee, J. W., Kim, E., Koo, J., Lee, K.** (2022). Representation selective self-distillation and wav2vec 2.0 feature exploration for spoof-aware speaker verification. *arXiv preprint arXiv:2204.02639*.

[58] **Teng, Z., Fu, Q., White, J., Powell, M. E., Schmidt, D. C.** (2022). SA-SASV: An end-to-end spoof-aggregated spoofing-aware speaker verification system. *arXiv preprint arXiv:2203.06517*.

[59] **Jung, J. W., Tak, H., Shim, H. J., Heo, H. S., Lee, B. J., Chung, S. W., … Kinnunen, T.** (2022). SASV 2022: The first spoofing-aware speaker verification challenge. *arXiv preprint arXiv:2203.14732*.

[60] **Zhang, Y., Li, Z., Wang, W., Zhang, P.** (2022). SASV based on pre-trained ASV system and integrated scoring module. *arXiv preprint arXiv:2207.00150*.

[61] **Wu, H., Kang, J., Meng, L., Zhang, Y., Wu, X., Wu, Z., … Meng, H.** (2022). Tackling spoofing-aware speaker verification with multi-model fusion. *arXiv preprint arXiv:2206.09131*.

[62] **Liu, X., Sahidullah, M., Kinnunen, T.** (2022). Spoofing-aware speaker verification with unsupervised domain adaptation. *arXiv preprint arXiv:2203.10992*.

[63] **Lin, J., Chen, T., Huang, J., Fang, R., Yin, J., Yin, Y., … Mao, Y.** (2022). The CLIPS system for 2022 spoofing-aware speaker verification challenge. *In INTERSPEECH (pp. 4367–4370)*.

[64] **Wang, X., Qin, X., Wang, Y., Xu, Y., Li, M.** (2022). The DKU-OPPO system for the 2022 spoofing-aware speaker verification challenge. *arXiv preprint arXiv:2207.07510*.

[65] **Martín-Doñas, J. M., Torre, I. G., Alvarez, A., Arellano, J.** (2022). The vicomtech spoofing-aware biometric system for the SASV challenge. *arXiv preprint arXiv:2204.01399*.

[66] **Mun, S. H., Shim, H. J., Tak, H., Wang, X., Liu, X., Sahidullah, M., … Jung, J. W.** (2023). Towards single integrated spoofing-aware speaker verification embeddings. *arXiv preprint arXiv:2305.19051*.

[67] **Heo, J., Kim, J. H., Shin, H. S.** (2022). Two methods for spoofing-aware speaker verification: Multi-layer perceptron score fusion model and integrated embedding projector. *arXiv preprint arXiv:2206.13807*.

[68] **Falez, P., Marteau, T.** (2024). Whispeak Speech Deepfake Detection Systems for the ASVspoof5 Challenge. In *Proc. ASVspoof 2024* (pp. 32-35).

[69] **Chen, Y., Wu, H., Jiang, N., Xia, X., Gu, Q., Hao, Y., ... Xu, M.** (2024). USTC-KXDIGIT System Description for ASVspoof5 Challenge. *arXiv preprint arXiv:2409.01695*.

[70] **Tran, T., Bui, T., Simatis, P.** (2024). ParallelChain Lab's Anti-spoofing Systems for ASVspoof 5. In *Proc. ASVspoof 2024* (pp. 9-15).

[71] **Duroselle, R., Boeffard, O., Courtois, A., Nourtel, H., Champion, P., Agnoli, H., Bonastre, J. F.** (2024, August). Data augmentations for audio deepfake detection for the ASVspoof5 closed condition. In *ASVspoof Workshop 2024* (pp. 16-23). ISCA.

[72] **Martín-Doñas, J. M., Roselló, E., Gomez, A. M., Álvarez, A., López-Espejo, I., Peinado, A. M.** (2024). ASASVIcomtech: the Vicomtech-UGR speech deepfake detection and SASV systems for the ASVspoof5 Challenge. *arXiv preprint arXiv:2408.10361*.

[73] **Villalba, J., Feng, T., Thebaud, T., Lee, J., Narayanan, S., Dehak, N.** (2024). The SHADOW Team Submission to the ASVSpoof 2024 Challenge. In *Proc. ASVspoof 2024* (pp. 36-42).

[74] **Aliyev, A., Kondratev, A.** (2024). Intema system description for the ASVspoof5 Challenge: power weighted score fusion. In *Proc. ASVspoof 2024* (pp. 152-157).

[75] **Rohdin, J., Zhang, L., Plchot, O., Staněk, V., Mihola, D., Peng, J., ... Burget, L.** (2024). BUT Systems and Analyses for the ASVspoof 5 Challenge. *arXiv preprint arXiv:2408.11152*.

[76] **Alenin, A., Balykin, A., Gómez, E., Makarov, R., Malov, P., Okhotnikov, A., ... Yakovlev, I.** (2024). IDVoice team System Description for ASVSpoof5 Challenge

[77] **Doddington, G. R., Przybocki, M. A., Martin, A. F., Reynolds, D. A.** (2000). The NIST speaker recognition evaluation–overview, methodology, systems, results, perspective. *Speech communication*, *31*(2-3), 225-254.

[78] **Kurnaz, O., Mishra, J., Kinnunen, T. H., & Hanilçi, C.** (2025). Optimizing a-dcf for spoofing-robust speaker verification. *IEEE Signal Processing Letters*.

[79] **Wang, X., Kinnunen, T., Lee, K. A., Noé, P. G., Yamagishi, J.** (2024). Revisiting and Improving Scoring Fusion for Spoofing-aware Speaker Verification Using Compositional Data Analysis. *arXiv preprint arXiv:2406.10836*.

[80] **Zhang, L., Li, Y., Zhao, H., Wang, Q., Xie, L.** (2022). Backend ensemble for speaker verification and spoofing countermeasure. *arXiv preprint arXiv:2207.01802*.

[81] **Mingote, V., Miguel, A., Ribas, D., Giménez, A. O., Lleida, E.** (2019). Optimization of False Acceptance/Rejection Rates and Decision Threshold for End-to-End Text-Dependent Speaker Verification Systems. In *INTERSPEECH* (pp. 2903-2907).

[82] **S. Chen *et al.*** (2022), "WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing," in *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505-1518, Oct. 2022, doi: 10.1109/JSTSP.2022.3188113.

[83] **Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., and Khudanpur, S.** (2018), "X-Vectors: Robust DNN Embeddings for Speaker Recognition," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, AB, Canada, 2018, pp. 5329-5333

[84] **Tak, H., Jung, J.-w., Patino, J., Kamble, M., Todisco, M., Evans, N**. (2021) End-to-end spectro-temporal graph attention networks for speaker verification anti-spoofing and speech deepfake detection. Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge, 1-8, doi: 10.21437/ASVSPOOF.2021-1

[85] **Wang, X., Delgado, H., Tak, H., Jung, J.-w., Shim, H.-j., Todisco, M., Kukanov, I., Liu, X., Sahidullah, M., Kinnunen, T.H., Evans, N., Lee, K.A., Yamagishi, J.** (2024) ASVspoof 5: crowdsourced speech data, deepfakes, and adversarial attacks at scale. Proc. The Automatic Speaker Verification Spoofing Countermeasures Workshop (ASVspoof 2024), 1-8, doi: 10.21437/ASVspoof.2024-1

[86] **Shim, H.-j., Jung, J.-w., Kinnunen, T., Evans, N., Bonastre, J.-F., Lapidot, I.** (2024) a-DCF: an architecture agnostic metric with application to spoofing-robust speaker verification. Proc. The Speaker and Language Recognition Workshop (Odyssey 2024), 158-164, doi: 10.21437/odyssey.2024-23

[87] **Kenny, P., Boulianne, G., Dumouchel, P.**, (2005), "Eigenvoice modeling with sparse training data," in *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 3, pp. 345-354, May 2005.

[88] **Kenny, P., Boulianne, G., Ouellet, P., Dumouchel, P.,** (2007) "Joint Factor Analysis Versus Eigenchannels in Speaker Recognition," in *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1435-1447.

[89] **Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., Ouellet, P.,** (2011) "Front-End Factor Analysis for Speaker Verification," in *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788-798, May 2011.

[90] **Ioffe, S.**, (2006). Probabilistic Linear Discriminant Analysis. In: Leonardis, A., Bischof, H., Pinz, A. (eds) Computer Vision – ECCV 2006. ECCV 2006. Lecture Notes in Computer Science, vol 3954. Springer, Berlin, Heidelberg. https://doi.org/10.1007/11744085_41.

[91] **Variani, E., Lei, X., McDermott, E., Moreno, I. L., Gonzalez-Dominguez, J.**, (2014) "Deep neural networks for small footprint text-dependent speaker verification," *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, 2014, pp. 4052-4056.

[92] **Snyder, D., Garcia-Romero, D., Povey, D., Khudanpur, S.** (2017) Deep Neural Network Embeddings for Text-Independent Speaker Verification. Proc. Interspeech 2017, 999-1003, doi: 10.21437/Interspeech.2017-620.

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

[94] **Cai, W., Chen, J., & Li, M.** (2018). Exploring the Encoding Layer and Loss Function in End-to-End Speaker and Language Recognition System. *ArXiv, abs/1804.05160*.

[95] **Zhou, T., Zhao, Y., Wu, J.,** (2021) "ResNeXt and Res2Net Structures for Speaker Verification," *2021 IEEE Spoken Language Technology Workshop (SLT)*, Shenzhen, China, 2021, pp. 301-307.

[96] **Roy, M. K., Keshwala, U.,** (2022) "Res2Net based Text Independent Speaker recognition system," *2022 12th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, Noida, India, 2022, pp. 612-616.

[97] **Yan, H., Lei, Z., Liu, C., Zhou, Y.,** (2024) "Gmm-Resnext: Combining Generative and Discriminative Models for Speaker Verification," *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seoul, Korea, Republic of, 2024, pp. 11706-11710.

[98] **Liu, B., Chen, Z., Wang, S., Wang, H., Han, B., Qian, Y**. (2022) DF-ResNet: Boosting Speaker Verification Performance with Depth-First Design. Proc. Interspeech 2022, 296-300, doi: 10.21437/Interspeech.2022-484.

[99] **Liu, T., Lee, K. A., Wang, Q., Li, H.,** (2024) "Golden Gemini is All You Need: Finding the Sweet Spots for Speaker Verification," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 2324-2337, 2024.

[100] **Chen, Y., Zheng, S., Wang, H., Cheng, L., Chen, Q., Qi, J.** (2023) An Enhanced Res2Net with Local and Global Feature Fusion for Speaker

Verification. Proc. Interspeech 2023, 2228-2232, doi: 10.21437/Interspeech.2023-1294.

[101] ISO/IEC 30107-1:2023, "Information technology biometric presentation attack detection."

[102] **Wang, X., Yamagishi, J., Todisco, M., Delgado, H., Nautsch, A., Evans, N.W., Sahidullah, M., Vestman, V., Kinnunen, T.H., LEE, K., Juvela, L., Alku, P., Peng, Y., Hwang, H., Tsao, Y., Wang, H., Maguer, S.L., Becker, M., & Ling, Z.** (2019). ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech. *Comput. Speech Lang., 64*, 101114.

[103] **Wu, Z., Kinnunen, T., Evans, N., Yamagishi, J., Hanilçi, C., Sahidullah, M., Sizov, A.** (2015) ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge. Proc. Interspeech 2015, 2037-2041, doi: 10.21437/Interspeech.2015-462.

[104] **Delgado, H., Todisco, M., Sahidullah, M., Evans, N., Kinnunen, T., Lee, K.A., Yamagishi, J.** (2018) ASVspoof 2017 Version 2.0: meta-data analysis and baseline enhancements . Proc. The Speaker and Language Recognition Workshop (Odyssey 2018), 296-303, doi: 10.21437/Odyssey.2018-42.

[105] **Jelil, S., Das, R.K., Prasanna, S.R.M., Sinha, R.** (2017) Spoof Detection Using Source, Instantaneous Frequency and Cepstral Features. Interspeech 2017, Stockholm, İsveç, 20-24 Ağustos 2017, 22-26.

[106] **Lavrentyeva, G., Novoselov, S., Malykh, E., Kozlov, A., Oleg, K., Shchemelinin, V.** (2017). Audio Replay Attack Detection with Deep Learning Frameworks. Interspeech 2017, Stockholm, Sweden, 20-24 August 2017, 82-86.

[107] **Kamble, M. R., Patil, H. A.** (2019) Analysis of Reverberation via Teager Energy Features for Replay Spoof Speech Detection. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 2019, 2607-2611.

[108] **Das, R. K.,Yang, J., Li, H.** (2019) Long Range Acoustic and Deep Features Perspective on ASVspoof 2019. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Singapore, 1018-1025.

[109] **Hajipour, M., Akhaee, M. A., Toosi, R.** (2021) Listening to Sounds of Silence for Audio replay attack detection, 7th International Conference on Signal Processing and Intelligent Systems (ICSPIS), Tehran, Iran, 1-6.

[110] **Gupta, P., Patil, H. A.** (2022) Linear Frequency Residual Cepstral Features for Replay Spoof Detection on ASVSpoof 2019. 30th European Signal Processing Conference (EUSIPCO), Belgrad, Serbia, 29 August- 2 September 2022, 349-353.

[111] **He, R., Cheng, Y., Zheng, Z., Ji, X., Xu, W.** (2024) Fast and Lightweight Voice Replay Attack Detection via Time-Frequency Spectrum Difference. IEEE Internet of Things Journal, vol. 11, no. 18, pp. 29798-29810.

[112] **Zeinali, H., Stafylakis, T., Athanasopoulou, G., Rohdin, J., Gkinis, I., Burget, L., Černocký, J.** (2019). Detecting Spoofing Attacks Using VGG and SincNet: BUT-Omilia Submission to ASVspoof 2019 Challenge. Interspeech 2019, Graz, Austria, 15-19 Eylül 2019, 1073-1077.

[113] **Yamagishi, J., Wang, X., Todisco, M., Sahidullah, M., Patino, J., Nautsch, A., Liu, X., Lee, K.A., Kinnunen, T., Evans, N., Delgado, H.** (2021)

ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection. Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge, 47-54, doi: 10.21437/ASVSPOOF.2021-8.

[114] **Patil, T., Acharya, R., Patil, H. A., Guido, R. C.** (2022) Improving the potential of Enhanced Teager Energy Cepstral Coefficients (ETECC) for replay attack detection. In: Computer, Speech & Language vol. 72, 101281.

[115] **Lei, Y., Huo, X., Jiao, Y., Li, Y.K.** (2021) Deep Metric Learning for Replay Attack Detection. Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge, 16 Eylül 2021, 42-46.

[116] **Buddhi, W., Eliathamby, A., Vidhyasaharan, S., Julien, E., Haizhou, L., Ting, D.** (2023) DNN controlled adaptive front-end for replay attack detection systems. Speech Communication, Vol. 154, 102973.

[117] **Xiangyu, S., Yuhao, L., Li, W., Haorui, H., Hao, L., Lei, W., Zhizheng, W.** (2023) Audio compression-assisted feature extraction for voice replay attack detection. arXiv, 2310.05813.

[118] **João, M., Jahangir, A., Falk, T. H.** (2020) Generalized end-to-end detection of spoofing attacks to automatic speaker recognizers. Computer Speech & Language, Vol. 63, 101096.

[119] **Zarish, A., Ali; J., Khalid, M.** (2022) AEXANet: An End-to-End Deep Learning based Voice Anti-spoofing System. Workshop on Artificial Intelligence for Multimedia Forensics and Disinformation Detection, 10356298.

[120] **Das, R.K.** (2021) Known-unknown Data Augmentation Strategies for Detection of Logical Access, Physical Access and Speech Deepfake Attacks: ASVspoof 2021. Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge, 16 Eylül 2021, 29-36.

[121] **Sanchez, J.C., Peinado, A.M., Gomez, A.M.** (2024) Data augmentation techniques for Physical Access in voice anti-spoofing. Proc. IberSPEECH 2024, 1-5, doi: 10.21437/IberSPEECH.2024-1.

[122] **Kurnaz, O., Kinnunen, T.H., Hanilçi, C., (2025)** "Investigating the Potential of Multi-Stage Score Fusion in Spoofing-Aware Speaker Verification," in 33. IEEE Conference on Signal Processing and Communications Applications (SIU2025) (Accepted).

[123] **Kurnaz, O., Demirtaş, S.C., Büker, A., Mishra, J., Hanilçi, C.** (2024) Spoofing-robust speaker verification using parallel embedding fusion: BTU speech group's approach for ASVspoof5 Challenge. Proc. The Automatic Speaker Verification Spoofing Countermeasures Workshop (ASVspoof 2024), 138-143, doi: 10.21437/ASVspoof.2024-20.

[124] **Kinnunen, T., Lee, K.A., Delgado, H., Evans, N., Todisco, M., Sahidullah, M., Yamagishi, J., Reynolds, D.A.** (2018) t-DCF: a Detection Cost Function for the Tandem Assessment of Spoofing Countermeasures and Automatic Speaker Verification . Proc. The Speaker and Language Recognition Workshop (Odyssey 2018), 312-319, doi: 10.21437/Odyssey.2018-44.

[125] **Chung, J.S., Nagrani, A., Zisserman, A.** (2018) VoxCeleb2: Deep Speaker Recognition. Proc. Interspeech 2018, 1086-1090, doi: 10.21437/Interspeech.2018-1929.

ÖZGEÇMİŞ

TARANMIŞ
VESİKALIK
FOTOĞRAF

Ad-Soyad **:** Oğuzhan KURNAZ

Doğum Tarihi ve Yeri **:**

E-posta **:**

ÖĞRENİM DURUMU:

- **Lisans** **:** 2016, Erciyes Üniversitesi, Mühendislik Fakültesi, Mekatronik Mühendisliği
- **Yüksek Lisans** **:** 2018, Erzurum Teknik Üniversitesi, Elektrik-Elektronik Mühendisliği ABD, Elektrik-Elektronik Mühendisliği Yüksek Lisans Programı

MESLEKİ DENEYİM VE ÖDÜLLER:

- Mekatronik Mühendisi / Elmos Otomasyon / 2015-2016
- Mekatronik Mühendisi / Mekatronik Sistemler / 2017-2017
- Elektrik Mühendisi / Birleşim Elektrik / 2017 / 2018
- Araştırma Görevlisi / Bursa Teknik Üniversitesi / 2018-Devam ediyor

TEZDEN TÜRETİLEN ESERLER, SUNUMLAR VE PATENTLER:

- **Kurnaz, O., Demirtaş, S.C., Büker, A., Mishra, J., Hanilçi, C.** (2024) Spoofing-robust speaker verification using parallel embedding fusion: BTU speech group's approach for ASVspoof5 Challenge. Proc. The Automatic Speaker Verification Spoofing Countermeasures Workshop (ASVspoof 2024), 138-143, doi: 10.21437/ASVspoof.2024-20
- **Kurnaz, O., Mishra, J., Kinnunen, T.H., Hanilçi, C.,** (2025) "Optimizing a-DCF for Spoofing-Robust Speaker Verification," in *IEEE Signal Processing Letters*, vol. 32, pp. 1081-1085, 2025, doi: 10.1109/LSP.2025.3545290.
- **Kurnaz, O., Kinnunen, T.H., Hanilçi, C., (2025)** “Investigating the Potential of

Multi-Stage Score Fusion in Spoofing-Aware Speaker Verification," in 33. IEEE Conference on Signal Processing and Communications Applications (SIU2025) (Accepted).