www.dergipark.gov.tr
ISSN:2148-3736
El-Cezerî Fen ve Mühendislik Dergisi
Cilt: 7, No: 3, 2020 (970-982)

El-Cezerî Journal of Science and Engineering
Vol: 7, No: 3, 2020 (970-982)
DOI : 10.31202/ecjse.707179

ECJSE

Makale / Research Paper

Konuşmacının Yaş ve Cinsiyetine Göre Sınıflandırılmasında DVM Çekirdeğinin Etkisi

Ergün YÜCESOY[1*]

[1]Ordu Üniversitesi, Teknik Bilimler Meslek Yüksekokulu, 52200, Ordu/TÜRKİYE
[a]yucesoye@odu.edu.tr

Received/Geliş: 21.03.2020 **Accepted/Kabul:** 03.05.2020

Öz: Bu çalışmada kısa süreli telefon konuşmalarından konuşmacının yaş ve cinsiyet grubunun otomatik olarak belirlenmesi konusu ele alınmıştır. Mel Frekansı Kepstrum Katsayıları (MFKK) ve bu katsayılardan türetilen delta parametreleri öznitelik olarak kullanılırken yaş ve cinsiyet sınıflarının modellenmesinde Genel Arkaplan Modelinden (GAM) uyarlanarak oluşturulan Gauss Karışım Modelleri (GKM) kullanılmıştır. Her konuşma için oluşturulan GKM modelleri süpervektörlere dönüştürülmüş ve bir Destek Vektör Makinesine (DVM) uygulanarak konuşmacıların yaş ve cinsiyet gruplarına göre sınıflandırılmıştır. GKM bileşen sayısı 32 ile 512 arasında değiştirilirken, doğrusal, polinomiyal, radyal tabanlı (RBF) ve GKM-KL olmak üzere dört farklı çekirdek işlevi kullanılarak testler yapılmış ve elde edilen sonuçlara göre önerilen sistem için en uygun bileşen sayısına ve çekirdek işlevine karar verilmiştir. aGender veritabanı ile yapılan testlerde en yüksek sınıflandırma oranı 256 bileşenli GKM'lerin GKM-KL çekirdeği ile sınıflandırılması sonucunda % 60.95 olarak elde edilmiştir.

Anahtar Kelimeler: Yaş ve cinsiyet tanıma; konuşma işleme; gauss karışım modeli; destek vektör makineleri.

The Effect of SVM Kernel on the Classification of the Speaker by Age and Gender

Abstract: In this study, the issue of automatic determination of the age and gender group of a speaker from short-term telephone conversations is discussed. While Mel Frequency Cepstrum Coefficients (MFCC) and delta parameters derived from these coefficients are used as feature, Gaussian Mixture Models (GMM) created by adapting from the universal background model (UBM) are used in modeling age and gender classes. After transforming the GMM models created for each speech into supervectors, they are applied as an input to a Support Vector Machine (SVM) and speakers are classified according to age and gender group. While the number of GMM components is changed between 32 and 512, tests are carried out using four different kernel functions; linear, polynomial, radial basis (RBF) and GKM-KL, and according to the results, optimum component number and kernel function are determined for the proposed system. In tests conducted with the aGender database, the highest classification rate was obtained as 60.95% as a result of classification of 256-component GMMs with GMM-KL kernel.

Keywords: Age and gender recognition; speech processing; gauss mixture model; support vector machines.

1. Giriş

Konuşma sinyali, seslendirilen ifadenin dilsel içeriğinin yanı sıra konuşmacının kimliği, yaşı, cinsiyeti, sosyal grubu, coğrafi bölgesi, sağlık durumu, sigara alışkanlığı, boyu ve psikolojik durumu gibi birçok kişisel bilgilerini de içerir. Paralinguistik olarak adlandırılan bu tür bilgiler insanların birbirleriyle olan iletişiminde sıklıkla kullanılır. Örneğin bir telefon konuşmasında konuşmacının sesinden bu tür bilgilerini çıkarır ve hitap şeklimizi ona göre belirleriz. İnsanlar arasındaki iletişimde olduğu gibi insan bilgisayar etkileşiminde de bu tür bilgiler farklı amaçlar için kullanılabilir. Örneğin etkileşimli sesli yanıt sistemlerinde, otomatik yaş ve cinsiyet tanıma

Bu makaleye atıf yapmak için
Yücesoy E., "Konuşmacının Yaş ve Cinsiyetine Göre Sınıflandırılmasında DVM Çekirdeğinin Etkisi" El-Cezerî Fen ve Mühendislik Dergisi 2020, 7(3); 970-982
Yücesoy E., "*The Effect of SVM Kernel on the Classification of the Speaker by Age and Gender*" El-Cezerî Journal of Science and Engineering, 2020, 7(3); 970-982.
ORCID: [1]0000-0003-1707-384X

yapılarak operatör için bekleyen müşterilere yaş ve cinsiyetine uygun reklam veya müzik sunulabilir [1]. Yaşlı müşteriler daha yavaş konuşan müşteri temsilcisine yönlendirilebileceği gibi en uygun müşteri temsilcisinin belirlenmesinde otomatik diyalekt veya aksan tanıma da düşünülebilir. Böylece müşteri temsilcisi ile yapılan görüşmelerde yanlış anlaşılmalar azaltılarak müşteri memnuniyeti arttırılabilir. Ses teknolojilerinin kullanıldığı bir diğer alan da adli vakalardır. Telefon kayıtlarının delil olarak kullanıldığı adli vakalarda kişinin sesinden yaşı, cinsiyeti, aksanı vb. bilgileri tespit edilebilir ve suçlunun belirlenmesinde bu bilgilerden yararlanılabilir [2]. Cinsiyet tanıma, konuşmaya dayalı birçok sistemde ön işlem olarak da kullanılabilir. Cinsiyet tanımanın ön işlem olarak kullanıldığı sistemlerde konuşmacının cinsiyetine göre cinsiyet bağımlı modeller oluşturulabilir ve bu modellerin kullanımı ile performans artışı sağlanabilir [3].

Son yıllarda konuşmacının sesinden yaş ve cinsiyetinin tanınması konusunda birçok çalışma yapılmıştır. Bu çalışmaların bazılarında yaş ve cinsiyet bilgileri birlikte ele alınırken, bazılarında ise ayrı olarak değerlendirilmiştir [4–9]. Cinsiyet tanıma konusunda yapılan çalışmalarda genellikle yetişkin konuşmacılar kullanılarak konuşmacıların erkek ve kadın olarak iki sınıfa ayrılması amaçlanmıştır [7,9]. Diğer taraftan yetişkin ve çocuk konuşmacıların birlikte ele alındığı çalışmalarda yapılmıştır. Bu çalışmaların çoğunda çocukların cinsiyet ayrımı yapılmamış ve konuşmacılar çocuk, erkek ve kadın olarak üç sınıfa ayrılmıştır [10,11]. Cinsiyet tanıma konusunda olduğu gibi yaş tanıma konusunda da farklı durumlar söz konusudur. Bu durumların ilki olan yaş grubu tanımada konuşmacının genç, yetişkin, yaşlı gibi yaş grubunun belirlenmesi amaçlanırken, yaş regresyonu olarak isimlendirilen ikinci durumda ise konuşmacının yıl olarak tam yaşının belirlenmesi amaçlanmıştır [8,12]. Literatürde yaş ve cinsiyet tanıma konusunda birçok çalışma yapılmıştır. Bunlardan Kockmann ve arkadaşları tarafından yapılan çalışmada [13] ifade tabanlı akustik, prosodik ve ses kalitesi öznitelikleri ile çerçeve tabanlı özniteliklerin birlikte kullanımı önerilmiştir. Çalışmada GKM ve DVM'ye dayalı modellemeden sonra doğrusal Gauss arka uçları ve lojistik regresyon temelli birleşim uygulanarak yaş sınıflandırmada %9, cinsiyet sınıflandırmada ise %4.5 iyileşme sağlanmıştır. Li ve arkadaşları tarafından yapılan çalışmada [4] beş farklı yöntemin skor seviyeli birleşimine dayalı yeni bir yaş ve cinsiyet belirleme yaklaşımı önerilmiştir. Çalışmada beş alt sistemin farklı kombinasyonda birleşimleri incelenmiş ve en iyi sınıflandırma oranı yaş ve cinsiyet kategorisinde %51.2, yaş kategorisinde %54.6 ve cinsiyet kategorisinde %84.7 olarak tüm alt sistemlerin birleştirilmesi sonucunda elde edilmiştir. Bakır tarafından yapılan çalışmada [7] ise konuşma sinyallerinden çıkarılan MFKK öznitelikleri gizli Markov modeli, dinamik zaman bükme ve yapay sinir ağı gibi yöntemlerle sınıflandırılmıştır. Erkek ve kadın konuşmacılar ayrı ayrı incelenmiş ve en iyi sınıflandırma oranı kadınlar için %98, erkekler için ise %97 olarak Markov modeli ile sağlanmıştır. Grzybowska ve arkadaşları tarafından yapılan çalışmada [12] yaş belirleme için i-vektörlerin kullanımına dayalı yeni bir yöntem önerilmiştir. Önerilen yöntemin eğitim aşamasında her yaş sınıfına karşılık gelen i-vektörlerin ortalaması alınmış, daha sonra test i-vektörleri ile yaş sınıflarının i-vektörleri arasındaki kosinüs uzaklığı hesaplanarak konuşmacıların yaş sınıfına karar verilmiştir. Çalışmada aGender veri kümesi ile yapılan testlerde %62.9 doğruluğa ulaşıldığı ve bu oranın Interspeech 2010 yarışmasındaki en iyi sonuçtan %16.7 daha yüksek olduğu belirtilmiştir. Qawaqneh ve arkadaşlarının çalışmasında [5] ileri beslemeli derin sinir ağları, darboğaz öznitelik çıkarıcı olarak kullanılmış ve dönüştürülmüş MFKK'lar i-vektör ve derin sinir ağı temelli sınıflandırıcılara uygulanarak konuşmacıların yaş ve cinsiyet sınıflandırması yapılmıştır. Çalışmada i-vektör sınıflandırıcının başarısı %56.13, derin sinir ağı sınıflandırıcısının başarısı ise %58.98 olarak verilmiştir. Safavi ve arkadaşları tarafından yapılan çalışmada [6] ise çocuk konuşmalarından kişi, cinsiyet ve yaş grubu tanıma konusu ele alınarak GKM-GAM, GKM-DVM ve i-vektör temelli yaklaşımların performans karşılaştırılması yapılmıştır. Konuşmacı tanımada, beklenildiği gibi yaş arttıkça hata oranının azaldığı, cinsiyet ve yaş grubu tanımada ise yaş ile başarı arasında daha karmaşık bir ilişkinin olduğu görülmüştür. Yapılan testlerde en iyi cinsiyet sınıflandırma başarısı yaştan bağımsız GKM-DVM sistemi ile %79.18 olarak, en iyi yaş grubu belirleme başarısı ise cinsiyetten bağımsız i-vektör temelli sistem ile %83 olarak elde edilmiştir. Büyük ve arkadaşları [8] tarafından yapılan bir diğer çalışmada ise yaş

sınıflandırma için GKM süpervektörlerinin ileri beslemeli derin sinir ağının girişi olarak kullanımı önerilmiştir. Önerilen sistem Türkçe ve Almanca ses kayıtlarından oluşan çok dilli bir veri kümesi ile test edilmiş ve yapılan testlerde sistemin yaş sınıflandırma başarısı erkekler için %49.7, kadınlar için ise %59.9 olarak belirlenmiştir.

GKM süpervektörlerine dayalı DVM yaklaşımı başta konuşmacı tanıma ve doğrulama olmak üzere birçok çalışmada kullanılmaktadır. Bu yöntemde konuşma sinyalinden çıkarılan öznitelikler GKM süpervektörlerine dönüştürüldükten sonra bir DVM sınıflandırıcısına uygulanmakta ve gerçekleştirilen optimizasyon sonucunda giriş vektörlerinin sınıfına karar verilmektedir. Bu noktada giriş vektörlerinin yüksek boyutlu uzaya taşınmasında kullanılan çekirdek işlevleri, DVM öğrenimini etkileyen önemli bir bileşen olarak karşımıza çıkmaktadır. Yaş ve cinsiyet sınıflandırma konusunda farklı çekirdek işlevlerinin kullanımına dayalı çeşitli çalışmalar yapılmıştır. Ancak bu çalışmalarda genellikle tek bir çekirdek işlevi üzerinde durulmuş, farklı çekirdek işlevlerinin yaş ve cinsiyet sınıflandırma üzerindeki etkileri ayrıntılı olarak araştırılmamıştır. Ayrıca bu çalışmalarda DVM optimizasyonunun nasıl yapıldığı ve hangi değerlerin (cost ve/veya gamma) kullanıldığı konusunda detaylı bilgiler verilmemiştir. Bu çalışmada birçok alanda yaygın olarak kullanılan doğrusal, polinomiyal ve radyal temelli DVM çekirdekleri ile Campbell ve arkadaşları [14] tarafından önerilen KL diverjansına dayalı GKM çekirdeğinin yaş ve cinsiyet tanıma üzerindeki etkileri ayrıntılı olarak incelenmiştir. Her bir çekirdek için ayrı ayrı parametre optmizasyonu yapılmış, belirlenen optimum parametreler ve elde edilen sonuçlar tablo halinde sunulmuştur. Çalışmada ayrıca yaş ve cinsiyet modeli olarak kullanılan GKM'lerin bileşen sayısının başarı üzerindeki etkisi de araştırılmıştır. Bileşen sayısı 32 ile 512 arasında değiştirilerek testler yapılmış ve elde edilen sonuçlara göre, geliştirilen sınıflandırma sistemi için en uygun bileşen sayısına ve DVM çekirdeğine karar verilmiştir.

Çalışmanın kalan kısımlarının konu akışı şöyledir; çalışmada kullanılan veri kümesi ve sınıf tanımlamaları Bölüm 2'de, MFKK öznitelik çıkarma yöntemi Bölüm 3'te, GKM modeli ve GKM süpervektörleri Bölüm 4'de tanıtılmıştır. Bölüm 5'de destek vektör makineleri ile kullanılan çekirdek işlevleri sunulurken, çalışma Bölüm 6'da verilen deneysel çalışmalar ve Bölüm 7'de verilen sonuç ve öneriler bölümleriyle sonlandırılmıştır.

2. Kullanılan Veritabanı ile Yaş ve Cinsiyet Sınıfları

Çalışmada önerilen yaş ve cinsiyet sınıflandırma sisteminin geliştirilmesinde aGender veritabanı kullanılmıştır [15]. aGender veritabanı, yaşları 7 ile 80 arasında değişen 954 Alman konuşmacının kendi özel telefonlarından (cep telefonu ya da sabit telefon) bir ses portalını arayarak gerçekleştirdikleri 65241 telefon görüşmesine ait ses kayıtlarından oluşmaktadır. Konuşmacıların yaş ve cinsiyet gruplarının otomatik olarak belirlenmesinde kullanılmak üzere oluşturulan aGender veritabanında, konuşmacılar yaş ve cinsiyetlerine göre 7 sınıfta tanımlanmıştır (Tablo 1). Veritabanındaki konuşmacıların yaş ve cinsiyet sınıflarına göre dağılımı yaklaşık eşit olup, çocuk konuşmacıların cinsiyet dağılımı da dengelidir. Tipik bir ses kaydının uzunluğu yaklaşık 2 s'dir. Ancak uzunluğu 3 ile 6 s arasında değişen konuşmalar da vardır. Bu konuşmalarda sabit (önceden belirlenmiş komut kelimeleri, ay, gün adları vs.) ve değişken (ad, soyad, telefon numarası vs.) içerikli 18 ifade, her konuşmacı tarafından farklı sayıda oturumda (en fazla 6) seslendirilmiş ve 8000 Hz 16 bit PCM formatında kaydedilmiştir. Toplam 47 saatlik telefon konuşmalarını içeren aGender veritabanı eğitim, geliştirme ve test olmak üzere üç bölümden oluşmaktadır. Bu bölümlerden eğitim ve geliştirme bölümlerindeki kayıtların yaş ve cinsiyet sınıfları veritabanında mevcuttur. Ancak test bölümündeki kayıtların sınıf bilgileri veritabanında paylaşılmamıştır. Eğitim ve geliştirme bölümlerindeki veriler yaş ve cinsiyet sınıflandırma sistemlerinin geliştirilmesinde, test bölümündekiler ise geliştirilen sistemlerin performans değerlendirmesinde kullanılmaktadır. aGender veritabanının eğitim ve geliştirme bölümündeki konuşmacı ve konuşma sayıları, yaş ve cinsiyet gruplarıyla birlikte Tablo 1' de verilmiştir.

Tablo 1. aGender veritabanının eğitim ve geliştirme bölümlerindeki konuşmaların yaş ve cinsiyet sınıflarına göre dağılımı

Sınıf	Grup	Yaş	Cinsiyet	#Eğitim	#Geliştirme
1	Çocuk	7-14	X	68/4406	38/2396
2	Genç	15-24	Erkek	63/4638	36/2722
3	Genç	15-24	Kadın	55/4019	33/2170
4	Yetişkin	25-54	Erkek	69/4573	44/3361
5	Yetişkin	25-54	Kadın	66/4417	41/2512
6	Yaşlı	55-80	Erkek	72/4924	51/3561
7	Yaşlı	55-80	Kadın	78/5549	56/3826

3. Öznitelik Çıkarma

Konuşma, farklı seviyelerde meydana gelen dönüşümlerin sonucu olarak ortaya çıkan karmaşık bir sinyal olup seslendirilen sözcüklerin içerdiği bilginin dışında konuşmacının yaş, cinsiyet, psikolojik durumu gibi kişisel bilgilerini de içerir. Ancak bu bilgilerin tamamına her zaman ihtiyaç duyulmaz. Bu nedenle konu ile ilgili bilgilerin konuşma sinyalinden çıkarılarak sinyalin sınırlı sayıda parametre ile temsil edilmesi gerekir. Öznitelik çıkarma olarak isimlendirilen bu işlem, konuşmaya dayalı tanıma sistemlerinde ilk adım olarak uygulanır ve tanıma performansını önemli derecede etkiler. Literatürde, konuşma sinyalinden öznitelik çıkarmak için geliştirilmiş birçok yöntem vardır. Mel Frekansı Kepstrum Katsayıları (MFKK) 1980'lerde Davis ve Mermelstein [16] tarafından tanıtılmış ve o tarihten itibaren en gelişmiş (state-of-the-art) yöntem olarak birçok çalışmada kullanılmıştır. MFKK öznitelik çıkarma yöntemi ön işlem, çerçeveleme, pencereleme, Fourier dönüşümü, Mel frekans kaydırma ve ayrık kosinüs dönüşümü gibi çeşitli adımları içermektedir. Bu adımların detayları aşağıda verilmiştir;

- *Ön-vurgulama:* Ön vurgulama, insan sesinin üretimi sırasında bastırılan yüksek frekanslı bölümlerin dengelenmesi işlemidir. Bu amaçla örneklenen konuşma sinyaline Denklem (1) ile gösterilen filtreleme işlemi uygulanır ve sinyalin yüksek frekanslı bölümlerinin enerjisi arttırılır.

$$Y[n] = X[n] - \alpha X[n-1] \tag{1}$$

Burada $Y[n]$ ön vurgulama işleminin çıkışını, $X[n]$ giriş konuşma sinyalini, α ise filtre katsayısını gösterir ve genellikle 0.9 ile 1.0 arasında seçilir [17,18]. Bu çalışmada α değeri 0.97 olarak alınmıştır.

- *Çerçeveleme ve Pencereleme:* Konuşma sinyali zamanla yavaş değişen bir sinyaldir. Ancak yeterince kısa zaman aralıklarında incelendiğinde durağan akustik özelliklere sahip olduğu görülmektedir. Bu nedenle konuşma sinyalleri kısa zaman aralıklarına bölünerek analiz edilir. Kısa süreli spektral ölçümler genellikle 20 ms uzunluğundaki bir pencere her 10 ms'de bir kaydırılarak gerçekleştirilir. Analiz penceresinin 10 ms kaydırılması konuşma sinyalinin zamansal karakteristiklerinin izlenmesini sağlarken, 20 ms'lik pencere genişliği ise genellikle iyi bir spektral çözünürlük için yeterlidir. Çerçevelere bölünen sinyalin sınırlarında oluşan süreksizlik, sinyale bir pencere fonksiyonu uygulanarak azaltılır. Denklem (2) ile verilen Hamming penceresi, ses işleme çalışmalarında en yaygın kullanılan pencere fonksiyonu olup bu çalışmada da kullanılmıştır.

$$w(n) = 0.54 - 0.46\cos\left(\frac{2n\pi}{N-1}\right), n = 0, \dots, N-1 \tag{2}$$

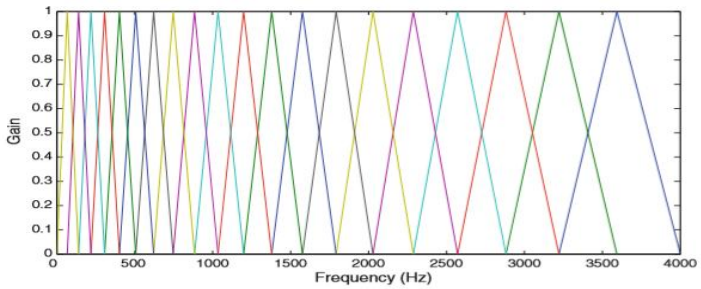
- *Ayrık Fourier Dönüşümü (DFT):* Pencereleme sonucunda oluşturulan her bir analiz çerçevesi üzerinde Denklem (3) ile verilen ayrık Fourier dönüşümü gerçekleştirilir ve sinyalin genlik spektrumu hesaplanır.

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-\frac{j2\pi kn}{N}}, 0 \leq k \leq N-1 \quad (3)$$

- *Mel spektrum:* Fourier dönüşümü uygulanan sinyal Mel-filtre kümesi olarak isimlendirilen bant geçiren bir filtre kümesinden geçirilerek sinyalin Mel spektrumu elde edilir. Mel, insan kulağının algıladığı frekansa dayanan bir ölçü birimidir. Bu birim doğrusal olarak sesin fiziksel frekansına karşılık gelmez. Çünkü insanın işitme sistemi perde frekanslarını doğrusal olarak algılamaz. Mel ölçeği yaklaşık olarak 1 kHz'in altında doğrusal üzerinde ise logaritmik frekans aralığına sahiptir [15]. Mel ile fiziksel frekans arasındaki ilişki yaklaşık olarak Denklem (4) ile ifade edilir.

$$f_{Mel} = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \quad (4)$$

Burada f Hz olarak fiziksel frekansı f_{Mel} ise algılanan frekansı temsil eder. Filtre kümeleri hem zaman uzayında hem de frekans uzayında gerçekleştirilebilir. MFKK hesaplamasında bu işlem genellikle frekans uzayında yapılır. Normalde filtrelerin merkez frekansları eşit aralıklıdır. Ancak insanın duyma algılamasını taklit etmek amacıyla filtre kümesinin frekans ekseni Denklem (4)'de verilen doğrusal olmayan bir fonksiyona göre kaydırılır. Yaygın olarak kullanılan filtre şekli üçgen olup Mel frekansına göre kaydırılmış bir üçgen filtre kümesi Şekil 1'de verilmiştir.



Şekil 1. Mel filtre kümesi

Sinyalin mel-spektrumu aşağıdaki ifadeye göre sinyalin genlik spektrumu ile Mel ölçekli üçgen filtrelerin her biri çarpılarak elde edilir.

$$s(m) = \sum_{k=0}^{N-1} [|X(k)|^2 H_m(k)], \quad 0 \leq m \leq M-1 \quad (5)$$

Burada M Mel ağırlıklı üçgen filtre sayısı olup genellikle 26 olarak seçilir [20]. $H_m(k)$ ise m'inci çıkış bandına katkıda bulunan k'ıncı enerji spektrum bölgesinin ağırlığıdır ve m 0 ile $M-1$ arasında olmak üzere aşağıdaki ifadeye göre hesaplanır.

$$H_m(k) = \begin{cases} 0, & k < f(m-1) \\ \frac{2(k-f(m-1))}{f(m)-f(m-1)}, & f(m-1) \leq k \leq f(m) \\ \frac{2(f(m+1)-k)}{f(m+1)-f(m)}, & f(m) < k \leq f(m+1) \\ 0, & k > f(m+1) \end{cases} \quad (6)$$

- *Ayrık Kosünüs Dönüşümü (DCT):* Bir önceki aşamada elde edilen Mel frekans katsayılarına ayrık kosinüs dönüşümü uygulanarak kepstral katsayılar üretilir. DCT işleminden önce genellikle Mel-Spektrumunun logaritması alınır. Logaritma işlemi ile genlik spektrumundaki çarpımsal bileşenler toplamsal bileşenlere dönüştürülür ve ters FFT'nin etkin bir versiyonu olan DCT ile bu bileşenler birbirlerinden ayrıştırılır. Bu işlem aşağıdaki ifade ile verilir [16].

$$c(n) = \sum_{m=0}^{M-1} \log_{10}(s(m)) cos\left(\frac{\pi n(m-0.5)}{M}\right), \; n = 0,1,2,\dots,C-1 \quad (7)$$

Burada $c(n)$ kepstral katsayıları, C ise MFKK sayısını gösterir. Sinyal bilgilerinin çoğu ilk birkaç MFKK katsayısı ile temsil edildiği için genellikle yüksek dereceli AKD bileşenleri yok sayılır ve ilk 8-13 kepstral katsayı öznitelik olarak kullanılır [22].

- *Dinamik MFKK öznitelikleri:* Kepstral katsayılar yalnızca belirli bir çerçevenin bilgisini içerdikleri için statik öznitelik olarak değerlendirilir. Sinyalin zamansal dinamikleriyle ilgili ilave bilgiler kepstral katsayıların birinci ve ikinci türevleri hesaplanarak elde edilir [23,24]. Birinci derece türev delta katsayıları olarak isimlendirilir ve konuşmanın hızı hakkında bilgi içerir. İkinci dereceden türev ise delta-delta katsayısı olarak isimlendirilir ve konuşmanın ivmesine benzer bilgiler sağlar. Dinamik parametrelerin hesaplanmasında yaygın olarak kullanılan ifade aşağıda verilmiştir [17].

$$\Delta c_m(n) = \frac{\sum_{i=-T}^{T} k_i c_m(n+i)}{\sum_{i=-T}^{T} |i|} \quad (8)$$

Burada $c_m(n)$ n'inci zaman çerçevesinin m'inci özniteliğini, k_i i'inci ağırlığı, T ise hesaplamada kullanılan ardışık çerçeve sayısını gösterir. Genellikle T 2 olarak seçilir. Delta-delta katsayıları ise delta katsayılarının birinci türevi alınarak hesaplanır.

4. Gauss Karışım Modeli (GKM)

Gauss karışım modeli, M bileşenli Gauss yoğunluklarının ağırlıklı toplamı olup aşağıdaki denklemle ifade edilir.

$$p(x|\lambda) = \sum_{i=1}^{M} w_i b_i(x|\mu_i, \Sigma_i) \quad (9)$$

Burada M bileşen sayısını, x D-boyutlu gözlem verisini, w_i $\sum_{i=1}^{m} w_i = 1$ şartını sağlayan bileşen ağırlıklarını, $b_i(x|\mu_i, \Sigma_i)$ ise ortalama vektörü μ_i ve kovaryans matrisi Σ_i ile karakterize edilen bileşen yoğunluklarını temsil eder. Her bileşen D-değişkenli bir Gaussian fonksiyonudur ve aşağıdaki ifade ile temsil edilir.

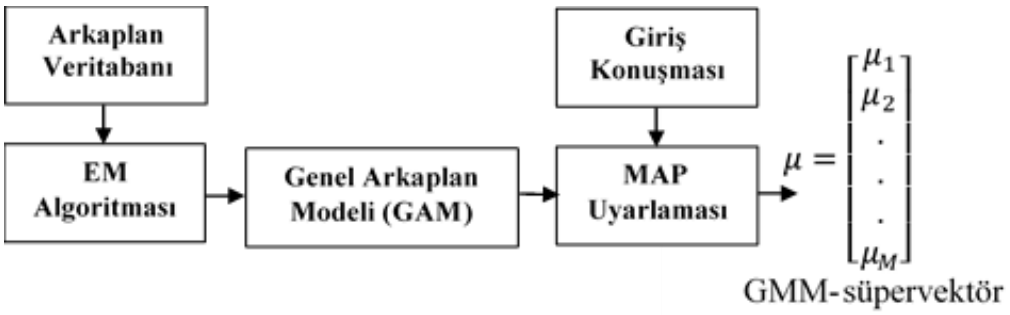
$$b_i(x|\mu_i, \Sigma_i) = \frac{1}{(2\pi)^{D/2}\sqrt{|\Sigma_i|}} exp\left\{-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)\right\} \quad (10)$$

Gauss karışım yoğunluğu, tüm bileşenlerin ağırlıkları, ortalama vektörleri ve kovaryans matrisleri ile aşağıdaki şekilde gösterilir. Bu parametrelerin tahmininde ise EM algoritması kullanılır [25].

$$\lambda = \{w_i, \mu_i, \Sigma_i\} \; i = 1, \dots, M \quad (11)$$

4.1. GKM Süpervektörleri

Konuşmacıların yaş, cinsiyet, psikolojik durum vb. özelliklerinin tanınması için konuşma sinyallerinden çıkarılan öznitelik vektörleri çeşitli sınıflandırıcılara uygulanır. Ancak her konuşmanın uzunluğu genellikle farklıdır ve bu konuşmalardan çıkarılan öznitelik vektörlerinin uzunlukları da farklı olacaktır. Diğer yandan çoğu sınıflandırıcı sabit uzunluklu vektörleri giriş olarak kabul eder. Bu nedenle değişken uzunluklu öznitelik vektörlerinin sabit uzunluklu vektörlere dönüştürülmesi gerekir. Blok diyagramı Şekil 2'de verilen GKM süpervektörleri bu fikirden ortaya çıkan bir yöntem olup işlem basamakları aşağıda verilmiştir.



Şekil 2. GKM-süpervektörlerin hesaplanması

İlk aşamada konuşmacıdan bağımsız temel model olarak işlev görecek olan Genel Arkaplan Modelinin (GAM) eğitimi için bir veritabanı belirlenir. Bu veritabanının tüm konuşmacı özelliklerini kapsayacak şekilde geniş bir veri kümesinden seçilmesi gerekmektedir. Ayrıca seçilen kayıtların konuşmacının yaşı, cinsiyeti, psikolojik durumu ve kayıt ortamı gibi farklı özelliklere göre dağılımlarının da eşit olması gerekmektedir. Aksi durumda oluşturulan GAM belirli bir gruba meyilli olacak, bu ise GAM'dan uyarlanarak oluşturulan sınıf bağımlı modellerin doğruluğunu azaltacaktır. Örneğin GAM'ın eğitiminde kullanılan veritabanında erkek konuşmacılar yoğunlukta ise bu veritabanı ile eğitilen GAM'da erkek grubuna meyilli olacaktır.

Genel arkaplan modelinin eğitiminde kullanılacak veritabanı belirlendikten sonra EM algoritması ile bir GKM eğitilir ve bu model GAM olarak kullanılır. Daha sonra her bir konuşma için GAM'ın parametreleri, ilgili konuşmanın taşıdığı bilgiyi temsil edecek şekilde uyarlanır. Uyarlama için genellikle MAP [26] yöntemi kullanılır ve GKM'lerin yalnızca ortalama vektörlerinin uyarlanması yaklaşımı tercih edilir. Böylece tüm GKM'lerin kovaryans matrisleri ayını olup, yalnızca ortalamaları farklı olacaktır. Her konuşmaya karşılık gelen GKM'nin GAM'dan uyarlanarak oluşturulmasının iki sebebi vardır. Bunlardan ilki, bir GKM'nin EM algoritması ile etkin bir şekilde eğitilmesi için büyük miktarda veriye ihtiyaç vardır. Ancak tek bir konuşmanın içeriği genellikle kısadır ve bu veri GKM'nin eğitimi için yeterli olmayabilir. GAM ise çok miktarda konuşmacı verisiyle eğitildiği için konuşmacıdan bağımsız öznitelikleri temsil edecektir. Bu nedenle her konuşmaya karşılık gelen GKM'lerin eğitiminde GAM, ön bilgi olarak kullanılabilir ve böylece GKM eğitimindeki veri eksikliği telafi edilebilir. İkinci sebep ise tüm GKM'ler GAM'dan yalnızca ortalama bileşenleri uyarlanarak oluşturulduğu için bu GKM'lerden elde edilen süpervektörler aynı uzayda karşılaştırılabilecektir. Son aşamada GAM'dan uyarlanarak oluşturulan GKM'lerin ortalama bileşenleri birleştirilerek GKM süpervektörleri oluşturulur. GKM süpervektörleri, konuşma ile yüksek boyutlu bir vektör arasındaki haritalama olarak düşünülür ve bu haritalama sonucunda tüm konuşmalar sabit boyutlu özniteliklere dönüştürülmüş olur.

5. Destek Vektör Makineleri

Destek Vektör Makineleri (DVM) Vapnik [27] tarafından tanıtılan ve konuşmaya dayalı tanıma sistemlerinde yaygın olarak kullanılan ikili (binary) bir sınıflandırıcıdır. İki sınıfı birbirinden ayıran aşırı düzlemin (hyperplane) konumu, en yakın eğitim vektörleri (destek vektörleri) ile arasındaki mesafe maksimum olacak şekilde seçilir. DVM doğrusal bir sınıflandırıcıdır, ancak Mercer koşullarını sağlayan çekirdek işlevlerinin kullanımı ile doğrusal olmayan sınırlara genişletilebilir. Bir DVM sınıflandırıcısı bir çekirdek işlevi $K(\cdot,\cdot)$'nın toplamı olarak aşağıdaki şekilde ifade edilir.

$$f(x) = \sum_{i=1}^{L} \alpha_i t_i K(x, x_i) + d \tag{12}$$

Burada $\sum_{i=1}^{L} \alpha_i t_i = 0$ ve $\alpha_i > 0$ olmak üzere t_i ideal çıkışları, d ise öğrenilmiş bir sabiti göstermektedir. x_i destek vektörleri olup bir optimizasyon süreci sonucunda eğitim kümesinden elde edilir [28]. İdeal çıkışlar, ilgili destek vektörün sıfırıncı veya birinci sınıfta olmasına göre 1 ya

da -1'dir. Sınıf kararı ise $f(x)$ değerinin belirli bir eşik seviyenin üzerinde veya altında olmasına bağlı olarak verilir. $K(\cdot,\cdot)$ çekirdeği Mercer koşullarını sağlayacak şekilde sınırlandırılarak aşağıdaki şekilde ifade edilebilir.

$$K(x,y) = b(x)^t b(y) \tag{13}$$

Buradaki $b(x)$ giriş uzayından olası sonsuz boyutlu bir genişleme uzayına haritalamayı temsil eder.

5.2. Uygulanan Çekirdekler

DVM bir çekirdek işlevi vasıtasıyla bir giriş uzayından yüksek boyutlu bir uzaya doğrusal olmayan bir haritalama gerçekleştirir. DVM eğitiminin en önemli bileşeni olan çekirdek için önerilmiş çeşitli alternatifler vardır. Bu çalışmada aşağıda verilen dört farklı çekirdek işlevi incelenerek bu çekirdeklerin önerilen sistemin performansına etkileri araştırılmıştır.

- Doğrusal çekirdek
- Polinomiyal çekirdek
- Radyal tabanlı (RBF) çekirdek
- GKM KL diverjans çekirdeği [14]

Bu çekirdeklerden ilk üçü DVM tabanlı sınıflandırmada yaygın olarak kullanılır ve sırasıyla denklem (14), (15) ve (16) ile temsil edilir.

$$K(x_i, x_j) = x_i^t x_j \tag{14}$$
$$K(x_i, x_j) = \left(x_i^t x_j + 1\right)^n \tag{15}$$
$$K(x_i, x_j) = exp\left[-\frac{1}{2}\left(\frac{\|x_i - x_j\|}{\sigma}\right)^2\right] \tag{16}$$

Burada n polinom derecesi, σ radyal tabanlı fonksiyonun genişliğidir. KL diverjans çekirdeği ise iki GKM arasındaki doğal uzaklığın bir yakınsaması olup μ^a ve μ^b ile verilen iki GKM süpervektörü için aşağıdaki şekilde tanımlanır.

$$\begin{aligned} K(utt_a, utt_b) &= \sum_{i=1}^{M} w_i \mu_i^a \Sigma_i^{-1} \mu_i^b \\ &= \sum_{i=1}^{M} \left(\sqrt{w_i}\Sigma_i^{-\frac{1}{2}}\mu_i^a\right)^t \left(\sqrt{w_i}\Sigma_i^{-\frac{1}{2}}\mu_i^b\right) \end{aligned} \tag{17}$$

Burada μ_i^a ve μ_i^b iki GKM'nin i'inci bileşenlerinin ortalamalarını, w_i ve Σ_i GAM'in i'inci Gauss bileşeninin ağırlık ve kovaryans matrisini (köşegen olduğu varsayılan) gösterir.

6. Deneysel Çalışmalar

Geliştirilen yaş ve cinsiyet sınıflandırma sistemi üzerinde aGender veritabanı kullanılarak testler yapılmıştır. Bu testlerde GKM bileşen sayısı 32 ile 512 arasında değiştirilirken DVM eğitiminde ise dört farklı çekirdek işlevi kullanılmıştır. Her bir bileşen sayısı ve çekirdek işlevi için sistem ayrı ayrı test edilmiş ve elde edilen sonuçlara göre en uygun çekirdek işlevine ve GKM bileşen sayısına karar verilmiştir. Çalışmada öznitelik olarak 13 MFKK katsayısı (sıfırıncı dahil) ile bu katsayıların birinci ve ikinci türevleriyle oluşturulan 39 elemanlı bir vektör kullanılmıştır. Öznitelik çıkarma aşamasında pencere genişliği olarak 25 ms, kayma miktarı olarak ise 10 ms seçilmiştir. Arkaplan modelinin eğitiminde aGender veritabanının eğitim bölümünden rastgele seçilen 70 konuşmacının yaklaşık 3.5 saatlik konuşmaları kullanılırken konuşmacıların seçiminde her yaş ve cinsiyet

sınıfından eşit sayıda (10'ar kişi) konuşmacının olmasına dikkat edilmiştir. GKM bileşen sayısı ikinin kuvvetleri olacak şekilde 32 ile 512 arasında değiştirilmiş ve yapılan testler sonucunda en uygun model büyüklüğüne karar verilmiştir. DVM'nin eğitiminde aGender veritabanının eğitim bölümündeki 331 konuşmacının 1571 oturumda seslendirdiği 22968 konuşma kullanılırken test aşamasında ise geliştirme bölümündeki 299 konuşmacının 1388 oturumda seslendirdiği 20489 konuşma kullanılmıştır. Çalışmada her oturumda seslendirilen konuşmalar bir bütün olarak değerlendirilmiş ve her oturum için bir GKM süpervektörü GAM'dan uyarlanarak oluşturulmuştur. Daha sonra bu vektörler DVM'ye giriş olarak uygulanmış ve gerçekleştirilen optimizasyon süreci sonunda her oturumdaki konuşmacıların yaş ve cinsiyet sınıfına karar verilmiştir.

Dört farklı DVM çekirdeği ve beş farklı GKM bileşen sayısı ile yapılan testlerin sonuçları Tablo 2'de verilmiştir. Bu sonuçlara göre konuşmacıların yaş ve cinsiyet gruplarına göre sınıflandırılmasında en yüksek başarı oranı %60.95 ile GKM-KL diverjansına dayalı çekirdeğin kullanımı sonucunda elde edilirken bu sonucu %59.87 ile 3 dereceli polinomiyal çekirdeği ve %59.73 ile RBF ve doğrusal çekirdekleri takip etmiştir. Yapılan deneylerde her çekirdek için en uygun parametrelerin (gamma ve cost) belirlenmesi amacıyla ızgara taraması yapılmıştır. Ceza parametresi -7 ile 15, gamma parametresi ise -17 ile 3 arasında taranmış ve elde edilen sonuçlara göre ilgili parametrelerin optimum değerlerine karar verilmiştir. Gerçekleştirilen parametre optimizasyonu sonucunda en yüksek başarının elde edildiği durumlar için belirlenen optimum DVM parametreleri Tablo 3'de verilmiştir.

Tablo 2. Farklı çekirdek ve bileşen sayıları ile geliştirilen GKM süpervektörlerine dayalı DVM sisteminin yaş ve cinsiyet sınıflandırma başarısı

GKM bileşen sayısı	GKM-KL	Doğrusal	Polinomiyal (3 dereceli)	RBF
512	60,45%	59,22%	59,22%	59,22%
256	***60,95%***	59,44%	59,87%	59,65%
128	60,52%	59,73%	59,73%	59,73%
64	59,01%	59,15%	58,93%	59,01%
32	57,93%	57,20%	57,64%	57,20%

Farklı bileşen sayıları ile yapılan testlerde ise kullanılan DVM çekirdeğinden bağımsız olarak GKM bileşen sayısının sınıflandırma başarısını arttırdığı ancak 256'dan sonraki artışın başarısı üzerinde önemli bir etkisinin olmadığı görülmüştür. GKM-KL çekirdeğinin ve polinomiyal çekirdeğinin kullanıldığı durumda en yüksek sınıflandırma başarısı 256 bileşenle sağlanırken, doğrusal ve RBF çekirdeklerinin kullanıldığı durumda ise en yüksek başarı 128 bileşenle sağlanmıştır. Elde edilen sonuçlar göz önünde bulundurulduğunda geliştirilen yaş ve cinsiyet sınıflandırma sistemi için en uygun çekirdek işlevinin GKM-KL çekirdeği, model büyüklüğünün ise 256 olduğu görülmüştür. Belirlenen optimum bileşen sayısı ve çekirdek işlevinin kullanıldığı durum için sistemin karışıklık matrisi Tablo 4'de verilmiştir.

Tablo 3. Izgara taraması sonucunda belirlenen optimum DVM parametreleri

Çekirdek işlevi	Ceza parametresi	Gamma parametresi
GKM-KL	1	-
Doğrusal	-7	-
Polinomiyal	8	-17
RBF	7	-14

Her sınıf için doğru ve yanlış öngörü oranlarını bir tablo şeklinde özetleyen karışıklık matrisi, sınıflandırma sistemlerinin performansını değerlendirmede sıklıkla kullanılan bir gösterim şeklidir.

Geliştirilen sınıflandırma sistemi için verilen karışıklık matrisi incelendiğinde; en iyi sınıflandırılan yaş ve cinsiyet grubunun genç kadın (GK) grubu olduğu, bu grubu sırasıyla yaşlı erkek (YaE), çocuk (Ç), yaşlı kadın (YaK), genç erkek (GE) ve yetişkin kadın (YeK) gruplarının takip ettiği, yetişkin erkek (YeE) grubunun ise en düşük sınıflandırma oranına sahip olduğu görülmektedir.

Tablo 4. Geliştirilen yaş ve cinsiyet sınıflandırma sisteminin karışıklık matrisi

	Ç	GK	GE	YeK	YeE	YaK	YaE
Ç	**66,05%**	12,35%	7,41%	6,17%	0,62%	7,41%	0,00%
GK	10,73%	**71,19%**	0,00%	14,69%	0,00%	3,39%	0,00%
GE	0,68%	2,04%	**60,54%**	2,04%	19,73%	2,72%	12,24%
YeK	2,23%	22,32%	0,00%	**54,91%**	0,45%	20,09%	0,00%
YeE	0,00%	0,00%	28,99%	1,78%	**40,83%**	0,00%	28,40%
YaK	0,82%	9,80%	0,00%	28,16%	0,00%	**60,82%**	0,41%
YaE	0,38%	0,00%	6,44%	0,38%	21,97%	1,52%	**69,32%**

Sınıflandırılma oranı en yüksek olan genç kadın grubunda test edilen 177 konuşmanın 19 tanesi çocuk (%10.73), 126 tanesi genç kadın (%71.19), 26 tanesi yetişkin kadın (%14.69) ve 6 tanesi de yaşlı kadın (%3.39) olarak sınıflandırılırken hiçbir konuşma erkek yaş gruplarına dahil edilmemiştir. Karışıklık matrisindeki hatalı kararların çoğunun aynı cinsiyetli konuşmacıların yaş grupları ve çocuk ile kadın yaş grupları arasında yoğunlaştığı görülmektedir. Erkek ve kadın grupları arasında en yüksek karışıklık genç erkek ile yaşlı kadın grubu arasında olurken, çocuk grubu ile en çok karışan grup genç kadın grubu olmuştur. Genç erkek ile yaşlı kadın ve çocuk ile genç kadın grupları arasında görülen yüksek karışık insan algılaması ile benzerlik göstermekte olup bu durum sonuçların dikkat çekici yönlerinden birisidir. Bir diğer dikkat çekici durum da yetişkin yaş grubunun sınıflandırma başarısının oldukça düşük olmasıdır. Geliştirilen sistemde genç kadınlar %71.19 doğrulukla sınıflandırılırken, yetişkin erkekler %40.83, yetişkin kadınlar ise %54.91 doğrulukla sınıflandırılmıştır. Bu durum yetişkin grubu ile genç ve yaşlı grupları arasında oluşan yüksek karışıklık ile ilişkili olup özellikle bu gruplar arasındaki karışıklığı azaltacak yeni özniteliklerin kullanımı ile sistemin performansında önemli bir artış sağlanabileceği şeklinde yorumlanmıştır.

Tablo 5. Önerilen sistemin performansının benzer çalışmalar ile karşılaştırılması

Çalışmalar	**Sınıflandırma Görevi**	**Veri kümesi**	**Doğruluk (%)**
Kockmann ve arkadaşları [13]	Gender		81.82
	Age	aGender	52.88
	Age & Gender		53.86
Li ve arkadaşları [4]	Gender		84.7
	Age	aGender	54.6
	Age & Gender		51.2
Bakır [7]	Gender	Belirtilmemiş	98 (kadın) 97 (erkek)
Grzybowska ve Kacprzak [12]	Yaş	aGender	62.9
Qawaqneh ve arkadaşları [5]	Age & Gender	aGender	58.98
Safavi ve arkadaşları [6]	Gender	OGI Kids Speech	79.18
	Age		83
Büyük ve Arslan [8]	Age	Multi-language	59.9 (kadın) 49.7 (erkek)
Bu çalışma	**Age & Gender**	**aGender**	**60.5**

Geliştirilen yaş ve cinsiyet sınıflandırma sisteminin sonuçları ile literatürdeki benzer çalışmaların sonuçları Tablo 5'de karşılaştırılmıştır. İncelenen çalışmaların çoğunda aGender veri kümesi kullanılmasına rağmen farklı veri kümeleri ile yapılmış çalışmalar da vardır. aGender veri kümesi konuşmacıların yaş ve cinsiyet gruplarına göre sınıflandırılmasında kullanılmak üzere özel olarak hazırlanmış bir veri kümesi olduğu için bu veri kümesi ile yapılan çalışmaların karşılaştırılması kolaydır. Diğer taraftan farklı veri kümeleri ile yapılan çalışmalarda kullanılan yaş ve cinsiyet sınıflarında bir standart yoktur ve bu çalışmaların sonuçlarının karşılaştırılması daha zordur. Örneğin bazı çalışmalarda iki (erkek, kadın), bazılarında ise üç sınıflı (çocuk, erkek ve kadın) cinsiyet sınıflandırma yapılmaktadır. Benzer şekilde yaş grubu sınıflandırmada da farklı durumlar söz konusudur. Konuşmacılar yaş grubuna göre iki sınıfa ayrılabileceği gibi üç veya dört sınıfa da ayrılabilmektedir. Ayrıca çalışmaların bir kısmında yaş ve cinsiyet sınıfları ayrı ayrı ele alınırken bazılarında ise birlikte değerlendirilmektedir. Bu bağlamda her çalışmada elde edilen sonuçların veri kümesindeki konuşmacı sayısı, konuşmacıların yaş ve cinsiyete göre dağılımı, kayıt ortamı ve sınıf sayıları gibi faktörler göz önünde bulundurularak değerlendirilmesi gerekmektedir.

Tablo 5'de verilen sonuçlar incelendiğinde, yaş ve cinsiyet sınıflarının birlikte ele alındığı çalışmalar arasında en yüksek başarının bu çalışmada elde edildiği görülmektedir. Ancak tüm çalışmalarda yaş ve cinsiyet sınıfları birlikte ele alınmamıştır ve bu durumda farklı görevlerde elde edilen sonuçların birbirleri ile kıyaslanması gerekmektedir. Farklı görevlerin kıyaslanmasında her üç görevin birlikte ele alındığı çalışmalarda elde edilen sonuçlar göz önünde bulundurularak yaklaşık bir değerlendirme yapılabilir. Verilen sonuçlardaki önemli bir detay da sınıf sayısıdır. aGender veri kümesi dışındaki veri kümeleri ile yapılan çalışmalarda cinsiyet sınıflandırmada iki, yaş sınıflandırmada ise üç sınıf kullanılmıştır. Oysaki aGender veri kümesinde cinsiyet sınıfı üç, yaş sınıfı ise dört sınıftan oluşmaktadır. Hem sınıf sayısı hem de gerçekleştirilen görevin zorluğu göz önünde bulundurulduğunda önerilen sistemin performansının incelenen tüm çalışmaların performansları ile rekabet edebilir seviyede olduğu görülmektedir.

7. Sonuç ve Öneriler

Bu çalışmada GKM süpervektörlerine dayalı DVM yaklaşımı ile konuşmacıların yaş ve cinsiyet gruplarına göre sınıflandırılması amaçlanmıştır. aGender veritabanı kullanılarak geliştirilen sistem dört farklı DVM çekirdeği ve beş farklı bileşen sayısı ile test edilmiş, elde edilen sonuçlara göre optimum model büyüklüğüne ve DVM çekirdeğine karar verilmiştir. Yapılan testler sonucunda önerilen yaş ve cinsiyet sınıflandırma sistemi için optimum bileşen sayısının 256, DVM çekirdeğinin ise GKM-KL çekirdeği olduğu görülmüştür. Belirlenen optimum bileşen sayısı ve DVM çekirdeğinin kullanıldığı durumda test edilen 1388 konuşmacının 846 tanesi doğru sınıflandırılarak %60.95 başarı sağlanırken üç dereceli polinomiyal çekirdek ile %59.87, doğrusal ve RBF çekirdekleriyle de %59.73 başarı sağlanmıştır. Yapılan testlerde hatalı kararların özellikle çocuk ve kadın yaş grupları arasında ve aynı cinsiyetli konuşmacıların yaş grupları arasında yoğunlaştığı, cinsiyet grubundan kaynaklanan hataların ise oldukça düşük olduğu görülmüştür. Her yaş ve cinsiyet grubunun sınıflandırılma oranı ayrı ayrı incelendiğinde ise bazı gruplar arasında büyük farkların olduğu saptanmıştır. Bu durum özellikle çocuk ile yetişkin grupları arasında ve aynı cinsiyetli konuşmacıların yaş grupları arasında ayırıcılığı daha yüksek özniteliklere ihtiyaç olduğunu ve birden fazla aşamada gerçekleştirilecek bir sınıflandırma yaklaşımı ile sınıflandırma performansında artış sağlanabileceği şeklinde yorumlanmıştır. Örneğin konuşmacılar önce çocuk ve yetişkin olarak, daha sonra yetişkin konuşmacılar erkek ve kadın olarak ve son aşamada erkek ve kadın konuşmacılar genç, yetişkin ve yaşlı olarak sınıflandırılabilir. Bu durumda her aşama için ilgili gruplar arasında ayırıcılığı daha yüksek farklı öznitelik kümelerin kullanımı mümkün olacaktır. Bu bağlamda MFKK gibi spektral özniteliklerin yanı sıra perde, süre ve yoğunluğa dayalı prosodik özniteliklerin ve HNR, jitter ve shimmer gibi ses kalitesine dayalı özniteliklerin birlikte ele alınabileceği düşünülmüştür.

Kaynaklar

[1]. Metze F., ve ark., "Comparison of Four Approaches to Age and Gender Recognition for Telephone Applications," 2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07, 15-20 April 2007, Honolulu, HI, USA, pp. IV-1089-IV-1092.

[2]. Tanner D. C., Tanner M. E., "Forensic aspects of speech patterns: voice prints, speaker profiling, lie and intoxication detection", Lawyers & Judges Publishing Company, 2004.

[3]. Bhukya S., "Effect of Gender on Improving Speech Recognition System", in International Journal of Computer Applications, 2018, 179(14), 22–30.

[4]. Li M., Jung C.-S., ve Han K., "Combining five acoustic level modeling methods for automatic speaker age and gender recognition", 11th Annual Conference of the International Speech Communication Association-INTERSPEECH 2010, 26-30 September 2010, Makuhari, Chiba, Japan, pp. 2826–2829.

[5]. Qawaqneh Z., Mallouh A. A., Barkana B. D., "Deep neural network framework and transformed MFCCs for speaker's age and gender classification", in Knowledge-Based Systems, 2017, 115, 5–14.

[6]. Safavi S., Russell M., Jančovič P., "Automatic speaker, age-group and gender identification from children's speech", in Computer Speech and Language, 2018, 50, 141–156.

[7]. Bakır C., "Automatic Speaker Gender Identification for the German Language", in Balkan Journal of Electrical and Computer Engineering, 2016, 4(2), 79–83.

[8]. Büyük O., Arslan L. M., "An investigation of multi-language age classification from voice", 12th International Conference on Bio-Inspired Systems and Signal Processing, BIOSIGNALS 2019 - Part of 12th International Joint Conference on Biomedical Engineering Systems and Technologies-BIOSTEC 2019, 22 - 24 February 2019, Prague, Czech Republic, pp. 85-92

[9]. Fokoue E., Ma Z., "Speaker Gender Recognition via MFCCs and SVMs", 2013, Accessed from https://scholarworks.rit.edu/article/1749

[10]. Přibil J., Přibilová A., Matoušek J., "GMM-based speaker age and gender classification in Czech and Slovak", in Journal of Electrical Engineering, 2017, 68(1), 3–12.

[11]. Faek F., "Objective Gender and Age Recognition from Speech Sentences", in Aro, The Scientific Journal of Koya University, 2015, 3(2), 24–29.

[12]. Grzybowska J., Kacprzak S., "Speaker Age Classification and Regression Using i-Vectors", in INTERSPEECH 2016, September 8–12 2016, San Francisco, USA, pp. 1402–1406.

[13]. Kockmann M., Burget L., ve Cernocký J., "Brno university of technology system for interspeech 2010 paralinguistic challenge", 11th Annual Conference of the International Speech Communication Association- INTERSPEECH 2010, 26-30 September 2010, Makuhari, Chiba, Japan, pp. 2822-2825

[14]. Campbell W. M., Sturim D. E., Reynolds D. A., ve Solomonoff A., "SVM based speaker verification using a GMM supervector kernel and NAP variability compensation", 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings, 14-19 May 2006, Toulouse, France, vol. 1, pp. 97–100.

[15]. Schuller B., ve ark., "The INTERSPEECH 2010 paralinguistic challenge", 11th Annual Conference of the International Speech Communication Association- INTERSPEECH 2010, 26-30 September 2010, Makuhari, Chiba, Japan, pp. 2794–2797.

[16]. Davis S. B., Mermelstein P., "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences", in IEEE Transactions on Acoustics, Speech, and Signal Processing, 1980, 28(4), 357–366.

[17]. Rabiner L., "Fundamentals of speech recognition", Pearson, 1 edition, 1993.

[18]. Hill P., "Audio and Speech Processing with MATLAB", CRC Press, 2018.

[19]. Stevens S. S., Volkmann J., Newman E. B., "A Scale for the Measurement of the Psychological Magnitude Pitch", in Journal of the Acoustical Society of America, 1937, 8(3), 185–190.

[20]. Kua J. M., ve ark., "Front-end diversity in fused speaker recognition systems", The Proceedings of APSIPA ASC 2010, 17-17 December 2010, Biopolis, Singapore, pp. 4-17.
[21]. Picone J. W., "Signal Modeling Techniques in Speech Recognition", in Proceedings of the IEEE, 1993, 81(9), 1215–1247.
[22]. Rao K. S., Vuppala A. K., "Speech processing in mobile environments", Springer International Publishing, 2014
[23]. Furui S., "Comparison of Speaker Recognition Methods Using Statistical Features and Dynamic Features", in IEEE Transactions on Acoustics, Speech, and Signal Processing, 1981, 29(3), 342–350.
[24]. Mason J. S., Zhang X., "Velocity and acceleration features in speaker recognition", 1991 International Conference on Acoustics, Speech, and Signal Processing-ICASSP 91, 14-17 April 1991, Toronto, Ontario, Canada, pp. 3673–3676.
[25]. Bilmes J., "A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models", International Computer Science Institute, Berkeley CA, 1998.
[26]. Azam M., Bouguila N., "Speaker verification using adapted bounded Gaussian mixture model", 2018 IEEE International Conference on Information Reuse and Integration (IRI), 6-9 July 2018, Salt Lake City, UT, USA, pp. 300–307.
[27]. Cortes C., Vapnik V., "Support-vector networks", in Machine learning, 1995, 20(3), 273–297.
[28]. Collobert R., Bengio S., "SVMTorch: Support Vector Machines for large-scale regression problems", in Journal of Machine Learning Research, 2001, 1(2), 143–160.