Avrupa Bilim ve Teknoloji Dergisi
Sayı 31 (Ek Sayı 1), S. 34-39, Aralık 2021

Araştırma Makalesi

www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
No. 31 (Supp. 1), pp. 34-39, December 2021

Research Article

Gerçek Zamanlı Ses Tanıma ile Robot Kolu Kontrolü

Ozan Fırat Çıplak[1], Serkan Keser[2*]

[1] Kırşehir Ahi Evran Üniversitesi, Fen Bilimleri Enstitüsü, Kırşehir, Türkiye, (ORCID: 0000-0002-5197-7810), ozanfc@gmail.com
[2*] Kırşehir Ahi Evran Üniversitesi, Müh.-Mim. Fakültesi, El.-Elektronik Müh. Bölümü, Kırşehir, Türkiye, (ORCID: 0000-0001-8435-0507), skeser@ahievran.edu.tr

(İlk Geliş Tarihi 10 Temmuz 2021 ve Kabul Tarihi 10 Aralık 2021)

(**DOI:** 10.31590/ejosat.969608)

ATIF/REFERENCE: Çıplak, O. F. & Keser, S. (2021). Gerçek Zamanlı Ses Tanıma ile Robot Kolu Kontrolü. *Avrupa Bilim ve Teknoloji Dergisi*, (31), 34-39.

Öz

Gün geçtikçe cihazların uzaktan kontrolünü gerçekleştiren tanıma sistemleri gelişmektedir. En çok kullanılan tanıma sistemleri olarak konuşma, yüz ve parmak izi tanıma sistemleri gösterilebilir. Konuşma tanıma sistemleri güvenlik sistemlerinde, cihaz kontrolü sistemlerinde ve dikte ettirme sistemlerinde gerçek zamanlı olarak kullanılabilmektedir. Bu çalışmada konuşma komutlarının gerçek zamanlı olarak tanınması ile robot kolu kontrolü gerçekleştirilmiştir. Konuşma komutlarının tanınması için Yapay Sinir Ağları (YSA), Fisher Doğrusal Ayrım Analizi (FDAA) ve Ayırt Edici Ortak Vektör (AOVY) sınıflandırıcıları kullanılmıştır. Eğitim kümesi için, her biri altı farklı renge sahip dört farklı nesne için toplam 24 adet konuşma cümleleri oluşturulmuştur. Eğitim kümesindeki konuşma sinyalleri 8 konuşmacı tarafından oluşturulmuştur. Test ve eğitim aşamalarında her kişi 50 konuşma sinyalli seslendirmiştir. Komutun tanınması ile robot kolu önceden koordinatları belli olan nesneye yöneltilmektedir. Çalışma sonucunda AOVY için dil modelli ortalama konuşma tanıma oranı %97,13 ve dil modelsiz %88,20 olarak bulunmuştur. FDAA için dil modelsiz ortalama konuşma tanıma oranı %87,3 ve dil modelli %96,3 olarak bulunmuştur. YSA için dil modelli ortalama konuşma tanıma oranı %89,76 ve dil modelsiz %82,3 bulunmuştur.

Anahtar Kelimeler: Konuşma tanıma, Robot kolu kontrolü, YSA, FDAA, AOVY.

Robot Arm Control with Real-Time Speech Recognition

Abstract

Recognition systems, which perform remote control of devices, are developing day by day. Speech, face, and fingerprint recognition systems seem to be the most frequently used recognition systems. Speech recognition systems can be used in real-time for security, device control and dictation systems. In this study, the robot arm is controlled by recognizing the real-time speech commands. Artificial Neural Networks (ANN), Fisher Linear Discrimination Analysis (FLDA) and Discriminative Common Vector Approach (DCVA) classifiers were used to recognize speech commands. For the training set, a total of 24 speech sentences have been recorded for four different objects with six different colors. Speech signals in the training set have been generated by 8 speakers. During the test and training phases, each person voiced 50 speech signals. The robot arm is directed to the objects whose coordinates are known beforehand with the recognition of the command. As a result of the study, the average speech recognition rate for DCVA with language model was % 97,13 and without language model was % 88,20. For the FLDA, the average speech recognition rate without language model was % 87,3 and with language model was % 96,3. For ANN, the average speech recognition rate with language model was % 89,76 and without language model % 82,3.

Keywords: Speech recognition, Robot arm controlling, ANN, FLDA, DCVA.

*Sorumlu Yazar: skeser@ahievran.edu.tr

1. Giriş

Ses tanıma sistemleri günümüzde pek çok değişik alanda kullanılmaktadır. Bu alanlar genel olarak sesli komutlar ile ev içi cihazların kontrolünün sağlandığı akıllı evler, robotların ve taşıtların kontrolü, interaktif sesli cevap sistemleri, sesin yazıya dönüştürüldüğü dikte ettirme, konuşmadaki duyguyu tanıma ve konuşmacı tanıma olarak sayılabilir (Filho ve Moir, 2010; Anggraeni, 2018; Soujanya ve Kumar, 2010; Furui ve ark., 2004; Beigi, 2011; Lalitha, 2015; Akyazi ve ark, 2019). Ses tanımayı etkileyen birçok etken bulunmaktadır. Bu etkenler genel olarak ses sinyallerine eklenen gürültüler, ses kaynağının ses alıcısına olan uzaklığı, seslendirilen kelimelerin yanlış telafuzu, seslendirilen sözcüklerin hangi sınıflayıcılar ile sınıflandırıldığı, kullanılan ses veri tabanı büyüklüğü, kişi bağımlı ya da kişi bağımsız olarak tanıma yapılması olarak sayılabilir (Çıplak ve Keser, 2021). Sınıflandırıcılar, ses tanıma oranlarının yüksek olmasında en büyük etkenlerden biridir. Sınıflandırıcıların tanıma oranlarına ise veri tabanı büyüklüğü ve kişi bağımlılık etkenleri de büyük etki etmektedir.

Literatürde en çok kullanılan ses tanıma sınıflayıcılarından biri olan Dinamik Zaman Bükme (DZB) algoritması, zaman serilerinin benzerlik ölçümünde kullanılan bir eşleştirme yöntemidir. Buna karşın genel tanıma oranı diğer sınıflayıcılara göre düşüktür (Permanasari ve ark., 2020). Saklı Markov Model (SMM) dil modeli kullanan, özellikle gerçek zamanlı ve kişi bağımsız tanımada yüksek tanıma oranları veren bir sınıflayıcıdır (Palaz ve ark.,2019; Muhammad ve ark.,2020; Tokuda ve ark., 2000). Uzun-Kısa Süreli Bellek (UKSB) mimarisini kullanan Tekrarlayan Sinir Ağları (TSA) ise günümüzde ses tanımada çok kullanılan ve iyi sonuçlar verdiği bilinen bir derin öğrenme algoritmasıdır (Sak ve ark. 2014; Dokuz ve Tüfekci, 2020). Ses tanımada kullanılan bir diğer derin öğrenme algoritması ise Evrişimsel Sinir Ağlarıdır (ESA). ESA algoritması sınıflayıcı olarak kullanılarak ses tanımada yüksek tanıma oranları elde edilebilmektedir (Guleti ve ark. 2020; Dokuz ve Tüfekci, 2020).

Ses tanımada kullanılan bir diğer sınıflayıcı ailesi alt uzay sınıflandırıcılardır. Literatürde görüntü yada ses tanımada kullanılan temel alt uzay sınıflayıcılar; Fisher Doğrusal Ayrım Analizi (FDAA), Class Featuring Information Compression (CLAFIC) ve Ortak Vektör Yaklaşımı (OVY) olarak sayılabilir (Keser ve Edizkan, 2009; Yavuz ve ark. 2006; Gunal ve Edizkan, 2008; Çıplak ve Keser, 2021). Bir alt uzay sınıflandırma yöntemi olan Ortak Vektör Yaklaşımı (OVY), sınırlı sayıda yalıtık kelime tanıma uygulamalarında yüksek tanıma oranları vermektedir (Gülmezoglu, 1999; Gunal ve Edizkan, 2008; Gülmezoğlu, 2007). Yapılan çalışmalarda sınırlı sayıda kelime kullanılarak %95'in üzerinde başarımlar elde edilmiştir. Ayrıca OVY (ya da AOVY) sınıflandırma yaparken yukarıda belirtilen pek çok sınıflayıcıya göre daha hızlı çalışabilmektedir. Hızlı çalışmasının temel sebebi ise her sınıfı temsil eden bir adet vektör kullanmasından kaynaklanmaktadır. Bu durum AOV'yi gerçek zamanlı ses tanıma uygulamaları için cazip hale getirmektedir. İlk olarak Çevikalp (2004) tarafından OVY yöntemini temel alan ve özellikle yüz tanıma uygulamalarında kullanılan Ayıredici Ortak Vektör Yaklaşımı (AOVY) tanıtılmıştır. AOVY yaklaşımı yüz tanıma uygulamalarında diğer alt uzay metotlar olan Eigenface ve FDAA sınıflandırıcılarına göre daha iyi sonuçlar verebilmektedir (Çevikalp, 2004). Ayrıca aynı AOV gibi, Eigenface ve FDAA alt uzay metotlara göre daha hızlı bir hesaplama süresine sahiptir. AOVY'nin OVY'ye göre bir avantajı ise test aşamasında her sınıf için sınıf sayısının bir eksiği boyutta öznitelik vektörleri ile işlem yapmasıdır. Genelde sınıf sayısı, OVY tarafından kullanılan öznitelik vektörü boyutundan çok daha küçük olmaktadır. Bir diğer önemli alt uzay yöntemi olan FDAA benzer şekilde genel olarak yüz tanıma çalışmalarında kullanılan ve Doğrusal Ayrım Analizi (DAA) yöntemini temel alan bir sınıflandırıcıdır (Belhumeur ve ark.,1997).

Bu çalışmada yalıtık kelime tanıma işlemi gerçek zamanlı ve kişi bağımlı olarak YSA, AOVY ve FDAA sınıflandırıcıları ile gerçekleştirilmiştir. Konuşma veri tabanı sınırlı sayıda (10 adet) sözcükten oluşmaktadır. Benzer bir çalışma her komut kelimesi için bir kişi tarafından seslendirilen 40 ses sinyali ile oluşturulmuştur (Çıplak ve Keser, 2021). Ancak kelime başına daha fazla ses sinyali ve daha fazla kişi ile eğitim yapma ile ortalama tanıma oranları daha sağlıklı yorumlanabilecektir. Buradan yola çıkarak eğitim ses veri tabanında bu sözcüklerin her biri için 50 ses sinyali 8 farklı kişi tarafından bilgisayar ortamında kaydedilmiştir. Ses sinyallerine MFCC uygulanarak öznitelik vektörleri elde edilmiştir. Eğitim aşamasında öncelikle ses veri tabanındaki her sözcüğe ait öznitelik vektörleri YSA, AOVY ve FDAA kullanılarak eğitilmiştir. Test aşamasında her kişinin bir mikrofon aracılığı ile seslendirdiği iki sözcüğün her biri ayrı ayrı bilgisayar ortamında yazılmış bir program aracılığıyla sınıflandırılmıştır. Komutun tanınması ile robot kolu önceden koordinatları belli olan nesneye yöneltilmiştir. Bu yönelme işleminin yapılabilmesi için öncelikle bilgisayar yazılım ara yüzü ile konuşma komutları gerçek zamanlı olarak tanınmakta ve tanınan komuta göre ilgili veriler RS232 seri iletişim protokolü kullanılarak robot kontrol kartına iletilmektedir. Ardından kontrol kartında her bir nesnenin yerinin bilgisini içeren mikrodenetleyici yardımı ile robotun servo motorları nesne konumuna doğru yönelmektedir. Sınıflama sonucunda elde edilen tanıma başarımını artırmak için bir çeşit dil modeli de geliştirilmiştir (Çıplak ve Keser, 2021). Dil modeli kullanılan sınıflandırma işleminde AOVY ile %97,13 tanıma oranına erişilirken, FDAA ile %96,3 ve YSA ile %89,76 tanıma oranına ulaşılmıştır.

2. Materyal ve Metot

2.1. Yalıtık Kelime Tanıma

Çalışmada ses tanıma için kullanılan yöntemler için öncelikle eğitim aşaması gerçekleştirilmiştir. Bunun için ilk olarak ses veri tabanı oluşturulmuş, ses veri tabanından her sınıf için öznitelik vektörlerinin elde edilmiş ve bu öznitelik vektörleri kullanılarak AOVY ve FDAA için eğitim işlemi yapılmıştır.

2.1.1. Eğitim Veri Tabanının Oluşturulması

Öncelikle 8 kişi tarafından kişi bağımlı olarak 10 farklı kelime seslendirilerek bilgisayar ortamına kaydedilmiştir. Bu 10 farklı kelimeden dördü "küp","prizma", "silindir", "küre" gibi şekilleri içerirken diğer altı kelime bu şekillere ait renklerden (beyaz, kırmızı, mavi, siyah, sarı, yeşil) oluşmaktadır. Robot kol tarafından ses tanıma ile algılanacak her bir şekle ait 6 renk oluşturulmuştur. Böylece toplam 24 farklı ikili sözcükten oluşan komut yapısı oluşmaktadır. Bu komutlar aşağıdaki tablo-1'de verilmiştir.

Tablo 1. Kullanılan iki kelimelik 24 adet komut kümesi

Renk-Nesne	Renk-Nesne	Renk-Nesne	Renk-Nesne	Renk-Nesne	Renk-Nesne
Kırmızı küp	Beyaz küp	Mavi küp	Siyah küp	Sarı küp	Yeşil küp
Kırmızı prizma	Beyaz prizma	Mavi prizma	Siyah prizma	Sarı prizma	Yeşil prizma
Kırmızı silindir	Beyaz silindir	Mavi silindir	Siyah silindir	Sarı silindir	Yeşil silindir
Kırmızı küre	Beyaz küre	Mavi küre	Siyah küre	Sarı küre	Yeşil küre

Çalışmada eğitim veri tabanı oluşturulurken nesnelere ve nesnelerin renklerine ait toplam 10 farklı kelime ("Küp"," Prizma",Silindir","Küre","Kırmızı","Beyaz","Mavi","Siyah","Sarı","Yeşil") 8 kişi tarafından mikrofon kullanılarak seslendirilmiştir. Eğitim ses veri tabanındaki her kelime 50 kez seslendirilmiştir. Böylece bir kelime sınıfı için toplam 400 (8×50) adet ses sinyali elde edilmektedir. Bu ses kayıtları her kelime sınıfı için 16 kHz'de örneklenmiştir ve her örnek 16 bit'liktir. Öz nitelik vektörleri 40 ms'lik çerçeveler üzerinden elde edilmiştir. Çerçeveler arası %50 üst üste bindirme yapılmıştır. Her bir ses sinyali için 32 çerçeve kullanılmıştır. Bu çerçevelerin her biri 13 MFCC katsayısı ve bir çerçeve enerjisinden oluşan 14 parametreyle temsil edilmektedir. Ayrıca delta ve delta-delta katsayıları da alınarak her çerçeve için toplam 42 özellik vektörü elde edilmiştir. Böylece 8 kişi için bir sınıfa ait her biri 1344 (n=1344) uzunluklu 400 (=50×8) adet öznitelik vektörü elde edilmiştir. Eğitim veri kümesi bu öznitelik vektörlerin her sınıf için birleştirilmesi ile oluşturulmuştur (Çıplak ve Keser, 2021). Toplam her sınıf için 400×8=3200 (m=3200) örnek oluşmakta ve bu durumda çalışmada AOVY için yeterli veri durumu (m≥n) oluşmaktadır.

2.2. Kullanılan Sınıflayıcılar

Çalışmada alt uzay sınıflayıcılar olan AOVY ve FDAA yanında YSA'da kullanılmıştır.

2.2.1. Ayırt Edici Ortak Vektör Yaklaşımı (AOVY)

Ortak vektör yaklaşımı (OVY) ile her sınıfa ait değişmez özellikleri taşıyan bir vektör elde edilir ve bu vektör "ortak vektör" olarak isimlendirilir (Gülmezoglu, 1999). AOVY ise elde edilen ortak vektörlerin birbirlerine göre dağılımlarını en büyükleyen dik iz düşüm vektör kümesi kullanmaktadır. Eğer eğitim setinde her biri k örnek olan c farklı sınıf varsa bu durumda eğitim setinde toplam *m=kc* adet örnek olacaktır. Burada *m* ses komut sınıfına ait vektör sayısını, *n* ise her bir vektör boyutunu göstermek üzere, OVY hem yeterli veri durumu ($\boldsymbol{m \geq n}$), hem de yetersiz veri durumlar ($\boldsymbol{m < n}$) için uygulanabilir (Keser ve Edizkan, 2009). Aynı durum AOVY içinde geçerlidir. Sınıfı i olan r'inci sinyal örneğini n-boyutlu uzayda $\boldsymbol{x_j^i}$ ile gösterirsek, sınıflar içi dağılım matrisi $\boldsymbol{S_w}$ aşağıdaki gibi verilir,

$$\boldsymbol{S_w} = \sum_{i=1}^{c} \sum_{r=1}^{k} \left((\boldsymbol{x_r^i} - \boldsymbol{\mu_i})(\boldsymbol{x_r^i} - \boldsymbol{\mu_i})^T \right) \qquad (1)$$

Burada, μ_i *i*'nci sınıfa ait ortalama vektörü göstermektedir. Farklılık alt uzayı **B** ve farksızlık alt uzayı $\mathbf{B}^{\perp}$ olmak üzere birbirine dik iki alt uzaya ayrılır (Gülmezoglu,1999). Farksızlık alt uzayı $\mathbf{B}^{\perp}$, $\boldsymbol{S_w}$ matrisinin sıfır öz değerlerine karşılık gelen öz vektörler tarafından gerilir. **P** ve $\overline{\mathbf{P}}$ matrisleri sırasıyla **B** ve $\mathbf{B}^{\perp}$ uzaylarının iz düşüm matrisleri olarak alınırsa, eğitim setindeki örneklerin $\mathbf{B}^{\perp}$ alt uzayındaki izdüşümleri aşağıdaki gibi olacaktır.

$$\boldsymbol{x_{com}^i} = \boldsymbol{x_r^i} - \mathbf{P}\boldsymbol{x_r^i} = \overline{\mathbf{P}}\boldsymbol{x_r^i}, \quad i=1,2,\ldots,c \qquad (2)$$

Aşağıdaki Denklem 3'te belirtilen $\boldsymbol{S_{com}}$ ortak vektörlere ait saçılım matrisi olup, aşağıdaki gibi bulunur,

$$\mathbf{S}_{com} = \sum_{i=1}^{C} (\boldsymbol{x_{com}^i} - \boldsymbol{\mu_{com}})\,(\boldsymbol{x_{com}^i} - \boldsymbol{\mu_{com}})^T. \qquad (3)$$

Bu eşitlikte $\boldsymbol{\mu_{com}}$ ortak vektörlere ait ortalama vektörünü ifade etmektedir. Bu durumda $\boldsymbol{S_{com}}$ matrisinin sıfırdan farklı öz değerlerine karşılık gelen öz vektörler, en uygun iz düşüm vektörlerini verir ($\mathbf{W_{opt}}$). Bu vektörlerin sayısı (r ≤ c-1), $\boldsymbol{S_{com}}$ matrisinin rankına eşittir. En uygun iz düşüm vektörleri üzerindeki iz düşüm katsayılarından oluşan öznitelik vektörleri aynı olup aşağıdaki gibi bulunur;

$$\boldsymbol{\Omega_i} = [< \mathbf{x_m^i}, \mathbf{w_i} > \cdots < \mathbf{x_m^i}, \mathbf{w_r} >] \qquad (4)$$

Bu vektörler, "ayırt edici ortak vektörler" olarak adlandırılır (Çevikalp, 2004). Test aşamasında ses sinyallerinin ayırt edebilmesi için öncelikle bu test sinyaline ait öznitelik vektörleri aşağıdaki eşitlikle bulunur:

$$\boldsymbol{\Omega_{test}} = \mathbf{W^T x_{test}} \qquad (5)$$

Daha sonra $\boldsymbol{\Omega_{test}}$ ile eğitim setindeki sınıflara ait ayırt edici ortak vektörlerin arasındaki Öklid uzaklığına bakılır. Test ses sinyali, en küçük uzaklığı veren sınıfa atanır. $\boldsymbol{\Omega_{test}}$ her sınıf için tek bir öznitelik vektörü ile karşılaştırıldığından tanıma oldukça hızlı gerçekleştirilebilmektedir.

2.1.2. Fisher Doğrusal Ayrım Analizi (FDAA)

FDAA ise DAA'dan türetilmiş bir alt uzay sınıflama metodudur. Bu metot kullanılarak sınıflar-arası ve sınıflar-içi dağılım oranını en büyükleyen bir dik vektör kümesi (**W**) bulunur (Belhumeur,1997). Burada sınıflar içi dağılım matrisi ($\boldsymbol{S_w}$) yukarıda belirtilen Denklem 1 ile bulunur. Sınıflar arası dağılım matrisi ise aşağıdaki gibidir,

$$\boldsymbol{S_B} = \sum_{i=1}^{C} N(\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})^T \qquad (6)$$

Burada *N* sınıfı oluşturan toplam örnek sayısıdır. $\boldsymbol{\mu_i}$ sınıf ortalaması ve $\boldsymbol{\mu}$ tüm sınıfların ortalamasıdır. En uygun taban vektörleri ($\mathbf{W_{opt}}$) aşağıdaki gibi bulunur (Belhumeur,1997).

$$\mathbf{W_{opt}} = \underset{\mathrm{W}}{\operatorname{argmax}} \frac{|W^T S_B W|}{|W^T S_W W|} = [\mathrm{w}_1\ \mathrm{w}_2\ \ldots\ \mathrm{w}_m] \qquad (7)$$

Burada *m=c*-1 olmaktadır. Aşağıdaki gibi belirtilen denklemde $\boldsymbol{S_W^{-1} S_B}$ çarpım sonucunda oluşan matrisin en büyük özdeğerlerine karşılık gelen *c-1* adet öz vektör en uygun taban vektörünü ($\mathbf{W_{opt}}$) vermektedir.

$$\boldsymbol{S_B w_i} = \boldsymbol{\lambda_i S_W w_i}, \quad i=1,2,\ldots,m \qquad (8)$$

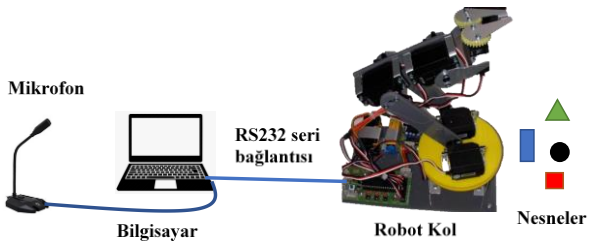
Tanıma problemlerinde, sınıf içi dağılım matrisi **Sw**'nin genellikle tekil olması problemiyle karşılaşılır. Bu problemin üstesinden gelmek için FDAA yöntemi ses sinyali kümesini daha düşük boyutlu bir uzaya iz düşürür. Bunun için öncelikle Temel Bileşen Analizi (TBA) kullanılarak özellik uzayı boyutu *N-c*'ye düşürülür. Ardından *N-c* boyutlu uzaya standart FLD uygulanarak boyut *c-1*'e düşürülür ve Denklem 7 uygulanarak $\mathbf{W}_{opt}$ matrisi bulunur. Eğitim kümesindeki her sınıfa ait öznitelik vektörleri $\mathbf{W}_{opt}$ kullanılarak optimum uzaya iz düşürülür. Test aşamasında, test sinyali $\mathbf{W}_{opt}$ kullanılarak iz düşürülüp $\mathbf{\Omega}_{test}$ bulunur. Ardından $\mathbf{\Omega}_{test}$ ile eğitim aşamasında sınıflara ait iz düşürülmüş vektörler arasındaki en küçük öklit mesafeyi veren sınıfa atama yapılır.

2.2.3. Yapay Sinir Ağları ile Sınıflandırma

Çalışmada eğitim aşaması için bir giriş katmanı, bir ara katman ve bir çıkış katmanı olmak üzere 3 katmanlı yapay sinir ağı modeli oluşturulmuştur. 1344 boyutlu öznitelik vektörleri çok büyük bir ağ yapısı gerektireceğinden öznitelik vektörlerine Temel Bileşen Analizi (TBA) uygulanarak boyutları 10, 20, 40, 60 ve 80 boyuta indirgenmiş ve her biri için test işlemi gerçekleştirilmiştir. Test verileri öncelikle gerçek zamanlı olarak elde edilip 1344 boyutlu MFCC katsayıları bulunmuştur. Ardından TBA uygulanarak boyutlar yukarıda belirtilen daha küçük boyutlara (10, 20, 40, 60 ve 80) indirgenerek ağın giriş katmanına uygulanmıştır. Çıkış katmanında elde edilen etiket değerine göre sınıflama gerçekleştirilmiştir.

3. Robot Kolu Kontrolü

Aşağıdaki Şekil 1'de gerçek zamanlı ses komutları tanınarak robot kolu kontrolü şeması verilmiştir.

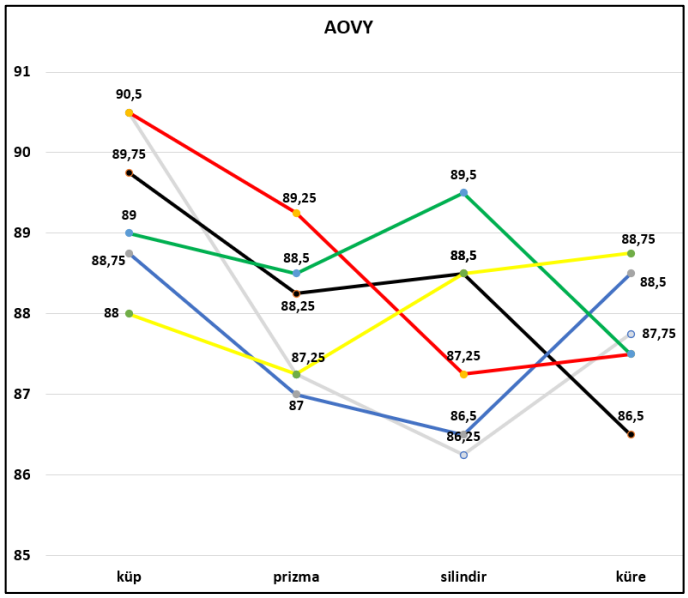


Şekil 1. Robot kolu kontrolü şeması

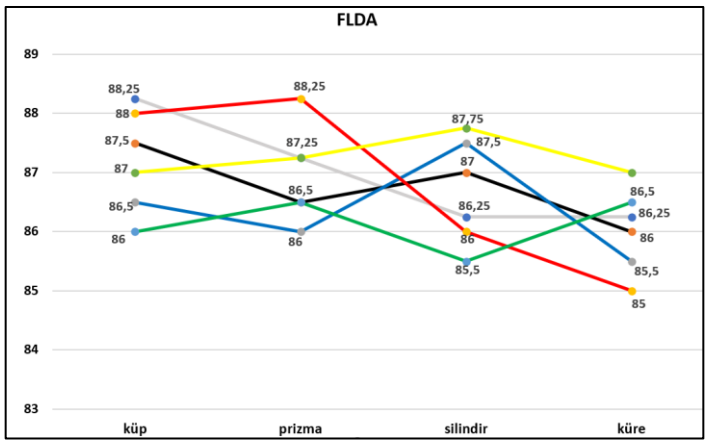
Test aşamasında ses sinyalleri 200-16000Hz frekans bant aralığında çalışan kapasitif bir mikrofon aracılığı ile gürültüsüz bir ortamda bilgisayara aktarılmakta ve komutları oluşturan her bir kelime sinyalinin başlangıç ve bitiş noktalarını hesaplayan bir algoritma yardımı ile bulunmaktadır. Sınırlar bulunduktan sonra test sinyali için MFCC ile öznitelik katsayıları elde edilmektedir. Ardından bu öznitelik vektörleri için AOVY ya da FDAA kullanılarak en olası sınıfa atama yapılmaktadır. Tanıma sonucunda bulunan komuta göre oluşturulan sinyal servo motor kontrol kartına iletilmektedir. Gelen sinyal tanınan komuta karşılık gelen nesnenin koordinat bilgisini içermektedir. Böylece robot kolu istenilen nesneye doğru yönelerek nesneyi kavramakta ve tanıma süreci son bulmaktadır. Başka bir komut geldiğinde benzer süreçler tekrarlanmaktadır.

4. Araştırma Sonuçları ve Tartışma

Kişi bağımlı kelime tanıma çalmasında her ikili kelime komutu 8 kişi tarafından 50 defa gerçek zamanlı olarak seslendirilerek test işlemi gerçekleştirilmiştir. Aşağıdaki şekillerdeki çubukların renkleri nesnelerin renklerine karşılık gelmektedir. Beyaz için gri renkli çubuk kullanılmıştır. Bu çalışmada AOVY için yeterli veri durumuna göre ses tanıma yapılmaktadır. Aşağıdaki Şekil 2'de ve Şekil 3'te dil modeli kullanmadan gerçekleştirilen testlerin sonuçları sırası ile AOVY ve FDAA için verilmiştir.



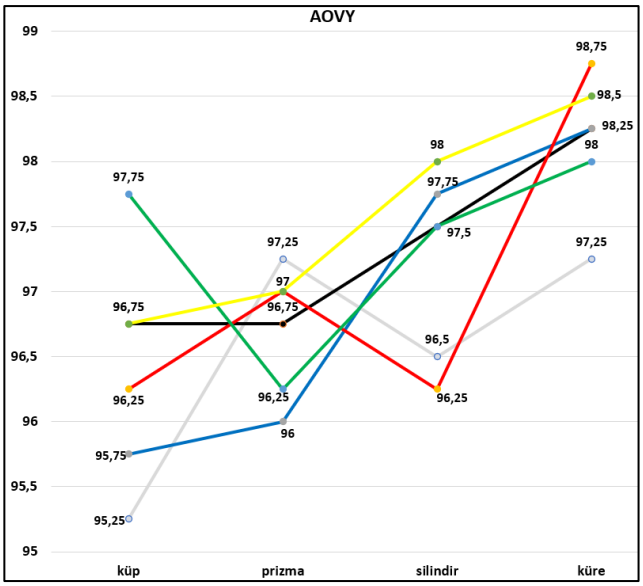
Şekil 2. AOVY için dil modelsiz komut kümeleri tanıma oranları (Ortalama=%88,20)



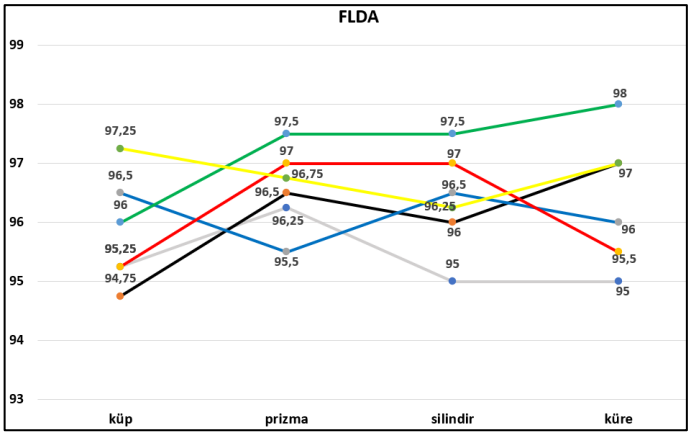
Şekil 3. FDAA için dil modelsiz komut kümeleri tanıma oranları (Ortalama=%86,72)

Şekil 2 ve Şekil 3'te dil modeli kullanılmadan AOVY ve FDAA için 24 farklı komut için bulunmuş ortalama tanıma oranları verilmiştir. 24 komutun ortalama tanıma oranları ise AOVY ve FDAA için sırası ile %88,2 ve %86,72 olarak bulunmuştur. Her ikili komutun kelimeleri ayrı ayrı 10 adet sınıf içerisinden en yakın sınıfa atanmıştır. Ancak bir komutta her zaman önce renk bilgisi sonra nesne bilgisi seslendirilmektedir. Bu yüzden ilk komutun 6 farklı renk içinden ve ikinci komutun 4 farklı nesne içinden seçilmesini içeren dil modelli bir çalışma daha gerçekleştirilmiştir. Bu şekilde her sözcük kendi sınıf kümesi içerisinden sınıflandırılacağı için daha iyi tanıma başarımları elde edilebilecektir. Aşağıdaki Şekil 4 ve Şekil 5'te dil modeli kullanılarak gerçekleştirilen testlerin sonuçları sırası ile AOVY ve FDAA için verilmiştir. 24 komutun ortalama tanıma oranları ise

AOVY ve FDAA için sırası ile %97,13 ve %96,3 olarak bulunmuştur.

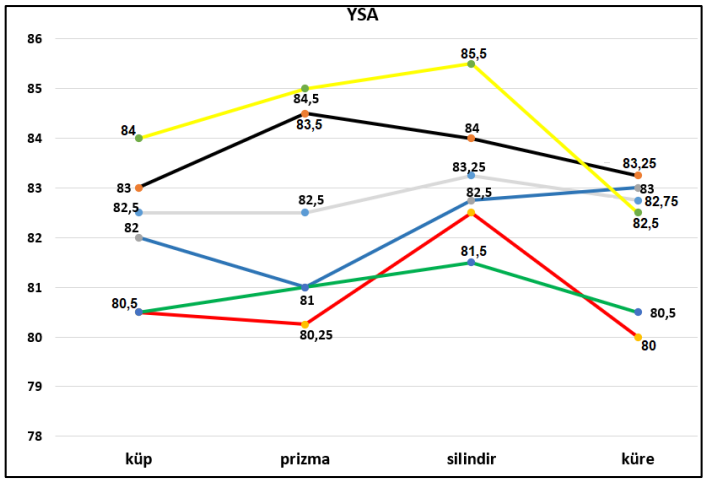


Şekil 4. AOVY için dil modelli komut kümeleri tanıma oranları (Ortalama= %97,13)

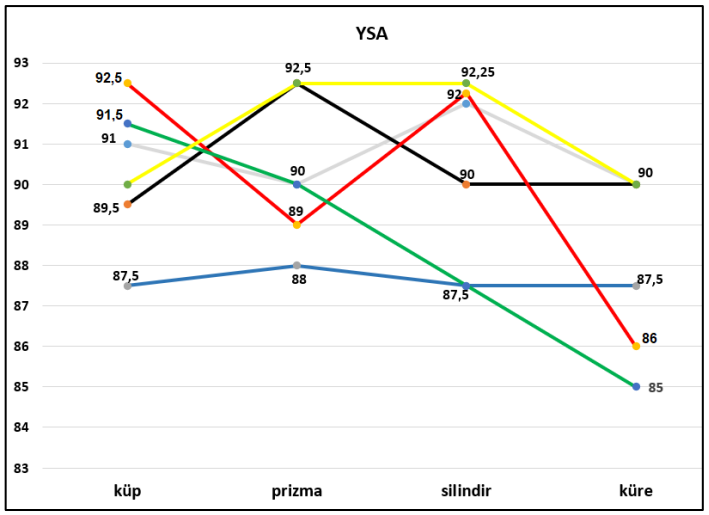


Şekil 5. FDAA için dil modelli komut kümeleri tanıma oranları (Ortalama= %96,3)

Şekil 4 ve Şekil 5'ten görülebileceği gibi dil modelli çalışma ile dil modelsiz çalışmaya kıyasla daha iyi tanıma oranları elde edilmiştir. Çalışmada TBA ile 1344 boyutlu MFCC öznitelik vektörleri sırası ile 10, 20, 40, 60 ve 80 boyuta indirgenmiş ve her biri için test işlemi gerçekleştirilmiştir. En iyi sonuç 40 boyutlu öznitelik boyutu için elde edilmiştir. Aşağıdaki şekiller için 40 boyutlu öznitelik vektörlerine göre 40 girişli bir giriş katmanı, ara katman ve bir çıkış katmanı olmak üzere 3 katmanlı yapay sinir ağı modeli oluşturulmuştur. Aşağıdaki Şekil 6 ve Şekil 7'de YSA için bulunan sırası ile dil modelsiz ve dil modelli sonuçlar verilmiştir. 24 komut için dil modelsiz ve dil modelli tanıma ortalama tanıma oranları sırası ile %82,42 ve %89,76 olarak bulunmuştur.



Şekil 6. YSA için dil modelsiz tanıma oranları (Ortalama=%82,42)



Şekil 7. YSA için dil modelli tanıma oranları (Ortalama=%89,76)

Çalışmada kullanılan sınıflayıcılar için bulunan ortalama tanıma oranları aşağıdaki Tablo 2'de verilmiştir.

Tablo 2. Sınıflayıcılar için bulunan ortalama tanıma oranları

Dil modelsiz (%)			Dil modelli (%)		
AOVY	**FDAA**	**YSA**	**AOVY**	**FDAA**	**YSA**
88,20	86,72	82,42	97,13	96,3	89,76

Tablo 2'den de görülebileceği gibi dil modelli tanıma oranları, dil modelsize göre daha yüksek çıkmıştır. Sınıflayıcılar arasından AOVY dil modelli ve dil modelsiz çalışmalarda en yüksek tanıma oranlarını verirken, FDAA ise AOVY'ye yakın değerler vermiştir. Ayrıca YSA, FDAA ve AOVY'den daha az tanıma oranları vermiştir. Sonuçlar incelendiğinde özellikle AOVY ve FLDA için gerçekleştirilen dil modelli çalışmaların, literatürdeki gerçek zamanlı ses tanımada yüksek tanıma oranlarına sahip SMM, TSA ve ESA sınıflayıcılar gibi başarılı sonuçlar verdiği görülmüştür.

5. Sonuç

Bu çalışmada sınıflandırma için etkin bir biçimde kullanılan alt uzay sınıflandırıcılar olan AOVY ve FDAA'nin yanı sıra YSA'da kullanarak üç farklı sınıflandırıcı ile gerçek zamanlı ve

kişi bağımlı konuşma tanıma uygulaması gerçekleştirilmiştir. AOVY ve FDAA genellikle görüntü tanıma uygulamalarında kullanılan sınıflayıcılardır. Ancak bu çalışma ile ses tanımadaki başarımları da test edilmiştir. Her komut 8 kişinin tanıma sonucu elde edilen bilgi seri haberleşme ile bilgisayara iletilmiş ve servo motor kontrolü sağlanarak robot kolu istenilen koordinatlara yönlendirilmiştir. Çalışma dil modelsiz ve dil modelli olarak iki biçimde gerçekleştirilmiştir. Dil modelsiz çalışmada 24 komut kümesinin tümü için AOVY ve FDAA sırası ile %88,2 ve %86,72 ortalama tanıma oranlarına erişmiştir. YSA ise dil modelsiz ortalama %82,42 tanıma oranına sahiptir. Dil modelli çalışmada ise AOVY ve FDAA için sırası ile %97,13 ve %96,3 ortalama tanıma oranlarına erişilmiştir. Dil modelli için YSA ise ortalama %89,76 tanıma oranına sahiptir. Sonuçlar incelendiğinde özellikle dil modelli tanımada AOVY ve FDAA'nin birbirine yakın ve oldukça iyi sonuçlar verdiği görülmektedir. YSA ise bu iki alt uzay sınıflayıcıdan daha düşük sonuçlar vermiştir. Ancak YSA önceden yapılan çalışmaya göre her sınıf için daha yüksek sayıda ses sinyali kullanıldığından daha iyi bir tanıma oranına erişildiği görülmüştür. AOVY ve FDAA için ise önceki çalışmanın tanıma oranlarına göre küçük bir miktar düşüş görülmüştür. Genel olarak değerlendirildiğinde özellikle dil modelli çalışmada AOVY ve FDAA'nın literatürde ses tanımada yüksek tanıma oranlarına sahip SMM, TSA ve ESA sınıflayıcıları gibi başarılı sonuçlar elde ettiği görülmüştür.

Kaynakça

Akyazi, Ö., Şahin, E., Özsoy, T., & Algül, M. (2019). A Solar Panel Cleaning Robot Design and Application. Avrupa Bilim ve Teknoloji Dergisi, 343-348.

Anggraeni, D., Sanjaya, W. S. M., Nurasyidiek, M. Y. S., & Munawwaroh, M. (2018). The implementation of speech recognition using mel-frequency cepstrum coefficients (MFCC) and support vector machine (SVM) method based on python to control robot arm. In IOP Conference Series: Materials Science and Engineering (Vol. 288, No. 1, p. 012042). IOP Publishing.

Beigi, H. (2011). Speaker recognition. In *Fundamentals of Speaker Recognition* (pp. 543-559). Springer, Boston, MA.

Belhumeur, P. N., Hespanha, J. P., & Kriegman, D. J. (1997). Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on pattern analysis and machine intelligence*, *19*(7), 711-720.

Çevikalp, H., Neamtu, M., Wilkes, M., & Barkana, A. (2004, April). A novel method for face recognition. In *Proceedings of the IEEE 12th Signal Processing and Communications Applications Conference, 2004.* (pp. 579-582). IEEE.

Çıplak, O.F., ve Keser S., (2021). Robot Arm Controlling With Real-time Speech Recognition Using Subspace Based Classifiers, *ISPEC 10th International Conference on Engineering & natural sciences*, (pp. 247-256), Siirt University, Siirt, TURKEY.

Dokuz, Y., & Tüfekci, Z. (2020). A Review on Deep Learning Architectures for Speech Recognition. Avrupa Bilim ve Teknoloji Dergisi, 169-176.

Filho, G. L., & Moir, T. J. (2010). From science fiction to science fact: a smart-house interface using speech technology and a photo-realistic avatar. *International journal of computer applications in technology*, *39*(1-3), 32-39.

Furui, S., Kikuchi, T., Shinnaka, Y., & Hori, C. (2004). Speech-to-text and speech-to-speech summarization of spontaneous speech. *IEEE Transactions on Speech and Audio Processing*, *12*(4), 401-408.

Gulati, A., Qin, J., Chiu, C. C., Parmar, N., Zhang, Y., Yu, J., ... & Pang, R. (2020). Conformer: Convolution-augmented transformer for speech recognition. arXiv preprint arXiv:2005.08100.

Gunal, S., & Edizkan, R. (2007, July). Use of novel feature extraction technique with subspace classifiers for speech recognition. In *IEEE International Conference on Pervasive Services* (pp. 80-83). IEEE.

Gunal, S., & Edizkan, R. (2008). Subspace based feature selection for pattern recognition. *Information Sciences*, *178*(19), 3716-3726.

Gülmezoglu, M. B., Dzhafarov, V., Keskin, M., & Barkana, A. (1999). A novel approach to isolated word recognition. *IEEE Transactions on Speech and Audio Processing*, *7*(6), 620-628.

Gülmezoğlu, M. B., Dzhafarov, V., Edizkan, R., & Barkana, A. (2007). The common vector approach and its comparison with other subspace methods in case of sufficient data. *Computer Speech & Language*, *21*(2), 266-281.

Keser, S., & Edizkan, R. (2009, April). Phonem-based isolated Turkish word recognition with subspace classifier. In *2009 IEEE 17th Signal Processing and Communications Applications Conference* (pp. 93-96). IEEE.

Lalitha, S., Mudupu, A., Nandyala, B. V., & Munagala, R. (2015, December). Speech emotion recognition using DWT. In *2015 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)* (pp. 1-4). IEEE.

Muhammad, H. Z., Nasrun, M., Setianingsih, C., & Murti, M. A. (2018, May). Speech recognition for English to Indonesian translator using hidden Markov model. In 2018 International Conference on Signals and Systems (ICSigSys) (pp. 255-260). IEEE.

Palaz, D., Magimai-Doss, M., & Collobert, R. (2019). End-to-end acoustic modeling using convolutional neural networks for HMM-based automatic speech recognition. Speech Communication, 108, 15-32.

Permanasari, Y., Harahap, E. H., & Ali, E. P. (2019, November). Speech recognition using dynamic time warping (DTW). In Journal of Physics: Conference Series (Vol. 1366, No. 1, p. 012091). IOP Publishing.

Sak, H., Senior, A., & Beaufays, F. (2014). Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition. arXiv preprint arXiv:1402.1128.

Soujanya, M., & Kumar, S. (2010, August). Personalized IVR system in contact center. In *2010 International Conference on Electronics and Information Engineering* (Vol. 1, pp. V1-453). IEEE.

Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., & Kitamura, T. (2000, June). Speech parameter generation algorithms for HMM-based speech synthesis. In *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 00CH37100)* (Vol. 3, pp. 1315-1318). IEEE.

Yavuz, H. S., Çevikalp, H., & Barkana, A. (2006). Two-dimensional CLAFIC methods for image recognition. In 2006 IEEE 14th Signal Processing and Communications