T.C.
KIRŞEHİR AHİ EVRAN ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
İLERİ TEKNOLOJİLER ANABİLİM DALI

# OVY, SVM, KNN ve TDNN SINIFLAYICILARI KULLANARAK KONUŞMACI BELİRLEME

**ESRA GEZER**

**YÜKSEK LİSANS TEZİ**

**KIRŞEHİR**
**2023**

T.C.
KIRŞEHİR AHİ EVRAN ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
İLERİ TEKNOLOJİLER ANABİLİM DALI

# OVY, SVM, KNN ve TDNN SINIFLAYICILARI KULLANARAK KONUŞMACI BELİRLEME

**ESRA GEZER**

**YÜKSEK LİSANS TEZİ**

**DANIŞMAN**
**Dr. Öğr. Üyesi Serkan KESER**

**KIRŞEHİR**
**2023**

**KIRŞEHİR AHİ EVRAN ÜNİVERSİTESİ FEN BİLİMLERİ ENSTİTÜSÜ**
**YÜKSEK LİSANS TEZ ÇALIŞMASI**
**ETİK BEYANI**

Kırşehir Ahi Evran Üniversitesi Bilimsel Araştırma ve Yayın Etiği Yönergesini okuduğumu ve anladığımı ve Kırşehir Ahi Evran Üniversitesi Fen Bilimleri Enstitüsü Tez Yazım Kurallarına uygun olarak hazırladığım bu tez çalışmasında;

- Tez içinde sunduğum verileri, bilgileri ve dokümanları akademik ve etik kurallar çerçevesinde elde ettiğimi,
- Tüm bilgi, belge, değerlendirme ve sonuçları bilimsel etik kurallarına uygun olarak sunduğumu,
- Tez çalışmasında yararlandığım eserlerin tümüne uygun atıfta bulunarak kaynak gösterdiğimi,
- Kullanılan verilerde ve ortaya çıkan sonuçlarda herhangi bir değişiklik yapmadığımı,
- Tez olarak sunduğum bu çalışmanın özgün olduğunu,

bildirir, aksi bir durumda bu konuda hakkımda yapılacak tüm yasal işlemleri ve aleyhime doğabilecek tüm hak kayıplarını kabullendiğimi beyan ederim. 25/08/2023

Öğrenci
Esra GEZER

**İÇİNDEKİLER DİZİNİ**

**Sayfa No**

**TEŞEKKÜR**

Yüksek lisans eğitimim boyunca engin bilimsel bilgisi ile her türlü desteği sağlayan, bilgi ve deneyimini benimle paylaşan, beni anlayış ve sabırla yönlendiren ayrıca bun tezin ortaya çıkmasını sağlayan tez danışmanım Dr. Öğr. Üyesi Serkan KESER'e minnet duyduğumu özellikle belirtmek isterim. Yüksek lisans eğitimim sırasında ailem her zaman yanımda olan ve beni teşvik eden aileme özel teşekkürlerimi iletmek isterim.

Ağustos, 2023 Esra GEZER

# ÖZET

## YÜKSEK LİSANS TEZİ

## OVY, SVM, KNN ve TDNN SINIFLAYICILARI KULLANARAK KONUŞMACI BELİRLEME

**Esra GEZER**

**KIRŞEHİR AHİ EVRAN ÜNİVERSİTESİ**
**FEN BİLİMLER ENSTİTÜSÜ**
**İLERİ TEKNOLOJİLER ANABİLİM DALI**

**Danışman:** **Dr. Öğr. Üyesi Serkan KESER**
**Yıl: 2023, Sayfa: 39**
**Jüri:** **Doç. Dr. Şekip Esat HAYBER**
**Doç. Dr. Mehmet Fatih TEFEK**
**Dr. Öğr. Üyesi Serkan KESER**

Konuşmacı tanıma çalışmaları günümüzde pek çok alanda kullanılmaktadır. Özellikle güvenlik sistemlerinde bu konu daha da önem kazanmıştır. Oluşturulacak konuşma tanıma sistemlerinin yüksek tanıma oranlarına erişmesi gerekir. Konuşmacı tanıma konuşmacı belirleme ve konuşmacı doğrulama olmak üzere ikiye ayrılır. Bu çalışmada Türkçe METUBET ve İngilizce MNIST veri tabanları için konuşmacı belirleme gerçekleştirilmiştir. Konuşmacı belirleme için MFCC katsayıları ve perde frekansı değerleri birleştirilmiştir. METUBET veri tabanı için 40 kişi, NMNIST veri tabanı için ise 30 kişi kullanılmıştır. Çalışmada OVY, SVM, KNN ve TDNN sınıflayıcılar kullanılmıştır. Konuşmacı belirlemede METUBET için en yüksek konuşmacı belirleme oranı SVM-polinom kernel ile %97.75 ve MNIST için TDNN ile %96.14 bulunmuştur. METUBET için konuşmacı tanıma sonucu OVY ile %100 bulunmuştur.

**Anahtar Kelimeler:** OVY, KNN, METUBET, MNIST, SVM-polinom kernel, TDNN

**ABSTRACT**

**MSc THESIS**

**SPEAKER IDENTIFICATION USING CVA, SVM, KNN, and TDNN CLASSIFERS**

**Esra GEZER**

**KIRŞEHİR AHİ EVRAN UNIVERSITY**
**INSTITUTE OF NATURAL AND APPLIED SCIENCES**
**DEPARTMENT OF ADVANCED TECHNOLOGIES**

| | |
|---|---|
| **Supervisor:** | **Assist. Prof. Dr. Serkan KESER** |
| | **Year: 2023, Pages: 39** |
| **Juries:** | **Assoc. Prof. Dr. Şekip Esat HAYBER** |
| | **Assoc. Prof. Dr. Mehmet Fatih TEFEK** |
| | **Assist. Prof. Dr. Serkan KESER** |

Speaker recognition studies are used in many fields today. Especially in security systems, this issue has gained more importance. Speech recognition systems to be created must reach high recognition rates. Speaker recognition is divided into speaker identification and speaker verification. In this study, speaker identification was carried out for the Turkish METUBET and English MNIST databases. MFCC coefficients and pitch frequency values are combined for speaker identification. 40 speakers were used for the METUBET database and 30 speakers were used for the NMNIST database. The CVA, SVM, KNN and TDNN classifiers were used in the study. In speaker identification, the highest speaker identification rate for METUBET was found to be 97.75% with SVM-polynomial kernel and 96.14% with TDNN for MNIST. Speaker recognition result for METUBET was found to be 100% with OVY.

**Keywords**: CVA, KNN, METUBET, MNIST, SVM-polynomial kernel, TDNN

## TABLOLAR DİZİNİ

**Sayfa No**

## ŞEKİLLER DİZİNİ

**Sayfa No**

## SiMGELER VE KISALTMALAR DİZİNİ

| Simgeler | | Açıklama |
|---|---|---|
| $f$ | : | Gerçek grekans |
| $L$ | : | Pencere uzunluğu |
| $\alpha$ | : | Hamming pencere sabiti |
| $S_w$ | : | Sınıf içi kovaryans matrisi |
| $V$ | : | Farksızlık alt uzayı |
| $\theta$ | : | Eşik değeri |

| Kısaltmalar | | Açıklama |
|---|---|---|
| **UBM** | : | Evrensel Arka Plan Modeli (Universal Backround Model (UBM)) |
| **GMM** | : | Gauss Karışım Modellemesi (Gaussian Mixer Model (GMM)) |
| **TDNN** | : | Zaman Gecikmeli Sinir Ağı (Time Delay Neural Net (TDNN)) |
| **KNN** | : | K en yakın Komşu (K Nearest Neigbor) |
| **SVM** | : | Destek Vektör Makineleri (Support Vector Machine) |
| **OVY** | : | Ortak Vektör Yaklaşımı |
| **MFCC** | : | Mel-Frekans Kepstral Katsayıları (Mel Frequency Cepstral Coefficients) |
| **JFA** | : | Ortak Faktör Analizi (Joint Factor Analysis (JFA)) |
| **WCCN** | : | Kovaryans Normalizasyonu |
| **LDA** | : | Doğrusal Ayrım Analizi (Linear Discriminant Analysis) |
| **PLDA** | : | Olasılıksal Doğrusal Ayrım Analizi |
| **ANN** | : | Yapay Sinir Ağı (Artifical Neural Net) |
| **CNN** | : | Konvolüsyonal Sinir Ağı (Convolution Neural Net) |
| **SOM** | : | Öz Yapılandırma Haritası (Self Organizing Map) |

# 1. GİRİŞ

Ses tanıma alanı, gelişen teknolojiyle birlikte hızla büyüyen bir alandır ve temel olarak insan sesinin mikrofon aracılığıyla bilgisayar tarafından algılanarak tanınması işlemini içerir. Bu süreç, insan-bilgisayar iletişiminde önemli bir rol oynamaktadır çünkü artık insanlar klavye kullanmadan bilgisayarlarına komut vermek istemektedirler. Microsoft'un, konuşma tanıma teknolojisini 1950'lerin sonlarından beri üzerinde çalışılan bir alan olarak gördüğü bilinmektedir.

Ses tanıma, oldukça zorlu bir çalışma disiplinidir ve bir dizi farklı alt problemleri içerir. Bu alt problemler, konuşmacı tanımlama, konuşmacı doğrulama, konuşmacıya bağımsız tanıma sistemleri, konuşmacıya bağımlı tanıma sistemleri, ayrık kelime tanıma, anahtar kelime algılama ve sürekli konuşma tanıma sistemlerini içerir. Sesin analizi ve tanınması sürecinde ses verisinin metne dönüştürülmesi odak noktasıdır. Bu dönüşüm süreci örnekleme, nicemleme ve kodlama aşamalarından oluşur. Örnekleme, sayısal işaretin belirli anlardaki genlik değerlerini içerir. Nicemleme, örneklenmiş işareti belirli aralıklara bölmeyi ve düzenlemeyi ifade eder. Kodlama, kuantize edilmiş işaretin bir sayı sisteminde temsil edilmesidir.

Konuşma tanıma alanı, sınırlı çalışmalar ve gereksinimler nedeniyle büyük ilgi çekicilik taşımaktadır. Ticari bakış açısından, konuşma tanıma teknolojisi büyük bir pazar potansiyeline sahiptir. İnsan sesinin farklılığı ve çalışılan dilin geniş içeriği, bu alanda çalışmanın zorluğunu artırmaktadır. Ancak, bu alanda yapılacak çalışmalar ve yenilikler insanlığın acil ihtiyaçlarından biridir. Ses tanıma teknolojisinin gelişmeleri sayesinde işitme engelli bireyler daha iyi iletişim kurabilirken, canlı yayınlardaki konuşmalar anında yazıya çevrildiğinde işitme sorunu olan bireyler de anlayabilirler. Ses tanıma, güvenlik işlemlerinde parmak izi gibi kullanılabilecek güvenilir bir teknolojidir. Bu alanda özellikle konuşmacı doğrulama ve konuşmacı tanımlama disiplinleri önemlidir.

Konuşmacı tanıma, ses işaretinin içerdiği bilginin kullanılarak kimin konuştuğunun otomatik olarak belirlenmesi işlemidir. Konuşmacı tanıma, konuşmacı doğrulama ve konuşmacı belirleme olarak iki temel kategoride incelenebilir. Konuşmacı doğrulama, bir ses örneğinin iddia edilen kişiye ait olup olmadığının tespitini içerir. Öte yandan, konuşmacı belirleme, verilen ses örneğinin sistemde kayıtlı olan kişilerden hangisine ait olduğunun saptanması anlamına gelir (Doddington, G. R., 1985; Campbell, J. P., 1997).

Bu işlemler metne bağımsız veya metne bağımlı olarak gerçekleştirilebilir. Metne bağımsız konuşmacı tanıma sistemlerinde eğitim ve test aşamalarında farklı cümleler veya sözcükler kullanılırken, metne bağımlı sistemlerde hem eğitim hem de test sırasında aynı cümle veya sözcükler tercih edilir.

Konuşmacı tanıma sistemleri birçok alanda kullanılır. Örneğin son yıllarda, telefon bankacılığı, sesli arama, telefonla alışveriş, veri tabanı erişim servisleri, cep telefonlarının sesle kontrolü ve bilgisayarların uzaktan sesle kontrolü gibi alanlarda yaygın olarak kullanılmaktadır (Furui, S., 1997).

Ses işaretinin sürekli değişken olması ve çevresel faktörlerin (gürültü, hava şartları vb.) etkisi, konuşmacı tanıma işlemini diğer problemlere kıyasla daha zorlu bir hale getirir. Ses işaretinin sürekli değişken olması, işlemlerin kısa zaman aralıklarında (10-20 ms) gerçekleştirilmesini gerektirir ve bu da veri boyutlarını artırır (Rabiner, L. R. ve Juang, B. H., 1993). Büyük veri boyutları, sınıflandırıcıların çalışma sürelerini uzatabilir. Bu nedenle, bir konuşmacı tanıma sistemi için kullanılan sınıflandırma yöntemi, yüksek başarı oranıyla birlikte hızlı sonuçlar üretebilmelidir.

Bu çalışmada konuşmacı tanıma için metinden bağımsız ve metin bağımlı konuşmacı belirleme gerçekleştirilmiştir. Veri tabanı olarak Türkçe olan METUBET ve İngilizce olan MNIST kullanılmıştır. METUBET için 40 konuşmacı, MNIST için ise 30 konuşmacı kullanılmıştır. Konuşmacı tanıma için sözcüklere ait çerçevelerden elde edilen MFCC katsayıları birleştirilerek yüksek boyutlu vektörler yerine sadece bir çerçeveden elde edilen 14 ya da 40 boyutlu MFCC katsayıları kullanılmıştır.

Konuşmacı belirlemede sınıflayıcı olarak zaman gecikmeli sinir ağı (TDNN), KNN, OVY ve SVM'den faydalanılmıştır. Ayrıca bu çalışmada METUBET ses veri tabanı kullanılarak konuşmacı tanıma programı yazılmıştır. OVY literatürde genel olarak kelime tanıma için kullanılmıştır, ancak bu çalışmada konuşmacı belirlemede kullanılmıştır.

## 1.1. Amaç ve Kapsamlar

Konuşmacının kimliğini belirleme işlemi, ses işaretinin içerdiği bilgi kullanılarak otomatik olarak gerçekleştirilir. Konuşmacı tanıma uygulamalarına, işlem kimlik doğrulaması, ücretli dolandırıcılığı önleme, telefonla kredi kartı satın alma, telefonla aracılık (örneğin, hisse senedi ticareti), çağrı merkezleri için müşteri bilgileri, ses örneği eşleştirme sayılabilir. Bu tanıma işlemi konuşmacı tanıma olarak adlandırılır ve konuşmacı doğrulama ile konuşmacı belirleme olmak üzere iki ana kategoriye ayrılır. Konuşmacı doğrulama ve konuşmacı belirleme, konuşma işleme ve güvenlik alanında kullanılan iki farklı kavramdır. Konuşmacı doğrulama, verilen bir ses örneğinin belirtilen kişiye ait olup olmadığını tespit ederken, konuşmacı belirleme, ses örneğinin önceden kaydedilmiş kişiler arasından hangisine ait olduğunu belirler (Doddington, G. R., 1985; Campbell, J. P., 1997).

Daha geniş ifadelerle, konuşmacı belirleme, bir ses kaydında veya konuşma verisinde yer alan konuşmacıların kimliklerini tespit etmeyi amaçlar. Bu süreç, bir dizi ses özelliğini analiz ederek belirli bir konuşmacının kimliğini belirleme işlemidir. Örneğin, bir ses kaydında hangi kişinin konuştuğunu tespit etmek için kullanılır. Bu tür teknolojiler, ses tabanlı kimlik doğrulama veya yetkilendirme sistemlerinde kullanılabilir.

Konuşmacı doğrulama, belirli bir kişinin, daha önce kaydedilmiş ses örnekleri ile karşılaştırılarak tanınmasını amaçlar. Bu süreç, kullanıcının kimliğini doğrulamak için mevcut ses örneği ile daha önce kaydedilmiş ses örnekleri arasındaki benzerlikleri analiz eder. Konuşmacı doğrulama, ses tabanlı güvenlik sistemlerinde kullanılır. Örneğin, telefon bankacılığı veya kimlik doğrulama uygulamalarında, kullanıcıların seslerini kullanarak kimliklerini doğrulamak için kullanılabilir.

Özetle, konuşmacı belirleme, bir ses kaydındaki farklı konuşmacıları tespit etmeyi amaçlarken, konuşmacı doğrulama, belirli bir kişinin kimliğini doğrulamayı hedefler. Her ikisi de ses tabanlı analizler kullanarak gerçekleştirilir ve genellikle ses işleme teknolojileri ve makine öğrenimi algoritmaları kullanılarak desteklenir. Konuşmacı tanıma işlemleri ayrıca metinden bağımsız ve metne bağımlı olarak iki gruba ayrılabilir. Metinden bağımsız sistemlerde, eğitim ve test aşamalarında farklı cümleler veya sözcükler kullanılırken, metne bağımlı sistemlerde eğitim ve test sırasında aynı cümle veya sözcükler kullanılır.

Konuşmacı tanıma sistemleri, çeşitli alanlarda yaygın bir şekilde kullanılır. Ancak ses işaretinin zaman içinde değişken olması ve dış etkenlerin etkisi nedeniyle, konuşmacı tanıma işlemi, diğer problemlere göre daha zorlu bir örüntü tanıma sorunudur. Büyük veri boyutları ise sınıflandırma süreçlerini yavaşlatabilir. Bu nedenle, bir konuşmacı tanıma sisteminde kullanılan sınıflandırma yöntemi hem yüksek performans göstermeli hem de hızlı sonuçlar üretmelidir.

## 2. ÖNCEKİ ÇALIŞMALAR

Mevcut çalışmalarda, konuşmacı tanımlama için otomatik konuşmacı tanımlama hattı olarak adlandırılan SI (Speaker Identification) sistemlerinin tasarım ve uygulama süreçleri tanımlanmıştır. Bu süreçler konuşma verisi toplama, sinyal ön işleme ve segmentasyon, akustik özellik çıkarma, boyut azaltma, sınıflandırma modelinin oluşturulması ve öğrenme modellerinin değerlendirilmesini içerir. Başka bir önemli aşama, sinyal ön işlemeyi içeren ham konuşma sinyalinin temsilidir. Genellikle, ham konuşma verisi, öğrenme modelinin yanlış sınıflandırılmasına yol açan birçok tepe ve arka plan gürültüsü içerir. Gürültü azaltma (Siam, ve ark., 2019), sessizlik kaldırma (Jahangir ve ark., 2020), ön vurgulama (An ve ark., 2019a), spektrogram temsili (Wang ve ark., 2020) ve uç nokta tespiti (Zhang ve ark., 2020) gibi birçok yöntem otomatik konuşmacı tanımlama için konuşma verisi ön işlemede önerilmiştir.

Segmentasyon teknikleri, ayırt edici akustik özellikleri çıkarmak için N sayıda çerçeveye bölünmüş konuşma sinyali elde etmek için örtüşme veya sabit pencere boyutları kullanır. Pencere boyutları, konuşmacı tanımlamanın mobil tabanlı uygulamasında hesaplama süresini azaltmak için önemli bir rol oynar. Segmentasyon tekniği genellikle kayan pencere, etkinlik veya enerji temelli yöntemleri içerir (Shi ve ark., 2020). Özellik çıkarma ve özellik seçimi, doğruluğu artırmak, hesaplama süresini azaltmak ve sınıflandırma hatasını düşürmek için ilgili özellik kümesini elde eder. Akustik özelliklerin çıkarılması, sığ ve derin özellikler olarak sınıflandırılabilir. Sığ özellikler, zaman alanı (istatistiksel), frekans alanı ve toplu deneme yanılma ayrıştırma (EMD) özelliklerini (Wu ve Lin, 2009a, 2009b) içeren geleneksel el yapımı özelliklerin çıkarılmasını içerir. Bununla birlikte, sığ özellikler büyük ölçüde alan uzmanı bilgisine bağlıdır, büyük miktarda etiketli konuşma verisi gerektirir ve genelleştirilebilir olması zor olan boyut azaltma tekniklerini kullanır. Son yıllarda derin öğrenme (Tran ve Tsai, 2020; Wang ve ark., 2020) ile ham konuşma verisinden otomatik özellik çıkarılması da otomatik konuşmacı tanımlamada sınıflandırma performansını artırmak için önerilmiştir. Derin öğrenme teknikleri, düşük düzeyden yüksek düzeye kadar çeşitli sinir ağlarının katmanlarını kullanarak ham konuşma verisinden ayırt edici özellikler çıkarmak için verilerin yüksek düzeyde temsilini kullanır.

Derin otokodlayıcı, tekrarlayan sinir ağı ve evrişimli sinir ağı gibi derin öğrenme teknikleri, model tanımlama, doğal dil işleme (Sutskever ve ark., 2014) ve şimdi otomatik konuşmacı tanımlamada çok popüler tekniklerdir. Konuşma verisinden çıkarılan özellikler, makine öğrenimi algoritmaları ile birleştirilerek bir konuşmacı tanımlama modeli oluşturulur. Bu makine öğrenimi algoritmaları, Gaussian Karışım Modeli (Al-Rawahy ve ark., 2012a), Destek Vektör Makinesi (Faragallah, 2018), k En Yakın Komşu (Sardar ve Shirbahadurkar, 2018b) ve Yapay Sinir Ağı (Wu ve Tsai, 2011) gibi algoritmaları içerir. Derin öğrenme için hem özellik çıkarma hem de sınıflandırma model oluşturmak için eğitilir. Son olarak otomatik konuşmacı tanımlama sistemi hassasiyet, hatırlama ve doğruluk gibi çeşitli performans metrikleri kullanılarak değerlendirilir.

Son yıllarda otomatik konuşmacı tanıma için farklı el yapımı özelliklerin, çoklu sınıflandırıcıların ve derin öğrenme sistemlerinin birleştirilmesi genellikle önerilmiştir (Calza ve ark., 2020; Fierrez, ve ark., 2018; Jahangir ve ark., 2020).

Ayrıca konuşmacı tanıma alanında birçok inceleme ve anket makalesi yayınlanmıştır (Larcher ve ark., 2014; Lawson ve ark., 2011; Saquib ve ark., 2010). Bununla birlikte, bu inceleme makaleleri otomatik konuşmacı tanıma için geleneksel el yapımı özelliklere ve makine öğrenmesi sınıflandırıcılarına odaklanmaktadır. Son zamanlarda, özellik çıkarım yöntemleri üzerine incelemeler (Disken ve ark., 2017; Tirumala ve ark., 2017) sunulmuştur. Tirumala ve arkadaşları (2017), konuşmacı tanıma uygulamaları için özellik çıkarım yöntemlerinde çeşitli yöntemleri ve algoritmaları tanımlayarak karşılaştırmış ve analiz etmiştir. Ancak çalışmamız, makine öğrenmesi sınıflandırıcılarından daha fazla derin öğrenme tekniklerini içeren son araştırmaları da kapsamaktadır. Benzer şekilde, Disken ve arkadaşları (2017) gürültülü, kanal uyumsuzluğu ve diğer bozulmuş koşullar altında konuşmacı tanıma için özellik çıkarım yöntemlerini tartışmaktadır. Son olarak, Tirumala ve Shahamiri (2016) tarafından sunulan, otomatik konuşmacı tanıma için derin öğrenme algoritmaları ve otomatik özellik çıkarımını tartışan bir inceleme yapılmıştır. Bu çalışma, genel derin sinir ağı mimarisi ve konuşmacı tanıma süreçlerine odaklanmıştır.

Almaadeed ve arkadaşları (2015) tarafından yapılan bir çalışma, dalgacık dönüşüm analizi ve yapay sinir ağları (ANNs) kullanarak metinden bağımsız bir konuşmacı tanıma (SI) sistemi tasarladı ve uyguladı. Bu yaklaşım, sınıflandırma hızını ve doğruluğunu artırmayı amaçlamaktadır.

Başlangıçta, dalgacık dönüşümü (Mallat, 1999; Vetterli ve Kovaˇcevic, 1995) verilen konuşma sinyalini birkaç seviyede daha küçük sinyal gruplarına ayırmak için uygulandı ve konuşma sinyalinin her bileşenini farklı frekanslarda ve çözünürlüklerde analiz etti. Daha sonra tüm konuşma sinyalinden ayırt edici özellikler çıkarıldı. Özellik çıkarma için kullanılan yöntemler arasında WSBC, WPT, DWT ve MFCC bulunmaktadır. Önerilen yöntem, SI modelinin oluşturulması için birden fazla sinir ağının birleştirilmesini içeriyordu. GRID konuşma veri tabanında yapılan değerlendirme, önerilen modelin klasik GMM, PCA ve BPNN'yi hem tanıma doğruluğunda hem de zamanda geride bıraktığını gösterdi.

Ayrıca, Soleymanpour ve Marvi (2017) MFCC özellik vektörlerini maksimum benzerlikle tanımlayan yeni bir yaklaşım önerdi. Bu, SI modelinin oluşturulmasında ve karar sınırının tanımlanmasında kullanıldı. MFCC özellikleri, konuşma sinyalinin her karesinden bir özellik vektörü olarak çıkarıldı ve daha sonra özellik vektörlerini maksimum benzerlikle elde etmek için K-means kümeleme yöntemi kullanıldı. ELSDSR konuşma veri tabanında yapılan deneylerde, SIN sistem performansının doğruluğunda iyileşme sağlandı ve bir yapay sinir ağı (ANN) sınıflandırıcı olarak kullanıldı.

Ünlü sesleri konuşmada daha yüksek enerjiyle daha sık meydana geldiğinden, gürültülü koşullarda ayırt edici özellikleri çıkarmak için ünlü fonemlerden yararlanmak mümkündür. Sarma ve Sarma (2013) bir konuşmacı tanıma için ünlü seslerin konuşmacı tarafından söylenen kelimelerden bölünmesini sağlayan yeni bir yöntem sundu. Ünlü sesin bölünmesi, olasılıksal sinir ağı (PNN) ve öz-yapılandırma haritasının (SOM) bir kombinasyonuyla gerçekleştirildi. Daha sonra, bölünmüş ünlü ses, konuşmacıların konuştuğu ünlü fonemlerin özelliklerini yakalayarak hazırlanan bir LVQ tabanlı kod kitabı ile desenlerle eşleştirilerek SI için kullanıldı. Önerilen SOM tabanlı yaklaşım, DWT tabanlı yaklaşıma göre %7'lik bir doğruluk artışı elde etti.

Ayrıca, Daqrouq ve Tutunji (2015), dalgacık entropisi, formantlar ve sinir ağlarını kullanarak yeni bir konuşmacı tanıma tabanlı özellik çıkarma yöntemi önerdi. İlk olarak, konuşmacıların sinyallerinden yedi Shannon dalgacık entropi paketi ve beş formant özellik vektörü olarak çıkarıldı. Geleneksel SI yöntemlerinin aksine, önerilen yöntem özellikleri kelimelerden (veya cümlelerden) değil, ünlülerden elde edildi. Daha sonra, bu 12 özellik katsayısı, beslemeli ileri sinir ağına girdi olarak verildi. Sadece 12 özellik katsayısı kullanarak, önerilen yöntem %89.16 doğruluk elde etti ve hem MFCC-ANN hem de LPC-ANN'yi geride bıraktı.

Yakın tarihli bir çalışmada, Dhakal ve arkadaşları (2019), Gabor Filtresi (GF) özelliklerini, CNN ile elde edilen özelliklerle ve istatistiksel özelliklerle bir matris olarak birleştirerek, SI performansını artırmak için yeni bir mimari önerdiler. CNN, 5 × 5 filtre boyutlu iki evrişim katmanı, 2 × 2 filtre boyutlu iki havuzlama katmanı ve bir tam bağlantılı katmandan oluşuyordu. 28 × 28 piksel boyutundaki gri tonlu görüntüler CNN'e giriş olarak veriliyordu. Hibrit özellik seti, DNN, rastgele orman (RF) ve SVM olmak üzere üç sınıflandırıcı kullanılarak sınıflandırıldı. Rapor edilen deneysel sonuçlar, RF'nin ELSDSR veri kümesinde %94.87'lik bir doğruluk elde ederek diğer iki sınıflandırıcıyı geride bıraktığını gösterdi.

# 3. KONUŞMACI TANIMAYA GİRİŞ

Konuşmacı tanıma, bir konuşmacıyı sesini kullanarak tanımlama görevidir. Konuşmacı tanıma iki bölüme ayrılır. Bunlar konuşmacı belirleme ve konuşmacı doğrulamadır. Konuşmacı belirleme, bilinen bir konuşmacı ses grubundaki hangi sesin hangi konuşmacıyla en iyi eşleştiğini belirleme süreci iken, konuşmacı doğrulama, akustik örneklerini analiz ederek bir konuşmacının kimlik iddiasını kabul etme veya reddetme görevidir. Konuşmacı doğrulama sistemleri, yalnızca bir veya iki model arasında bir karşılaştırma gerektirdiğinden, konuşmacı belirleme sistemlerinden hesaplama açısından daha az karmaşıktır, oysa konuşmacı belirlemede, bir modelin N adetlik konuşmacı modelleriyle karşılaştırılmasını gerektirir (Çeliktaş H., 2019).

## 3.1. Konuşmacı Doğrulama

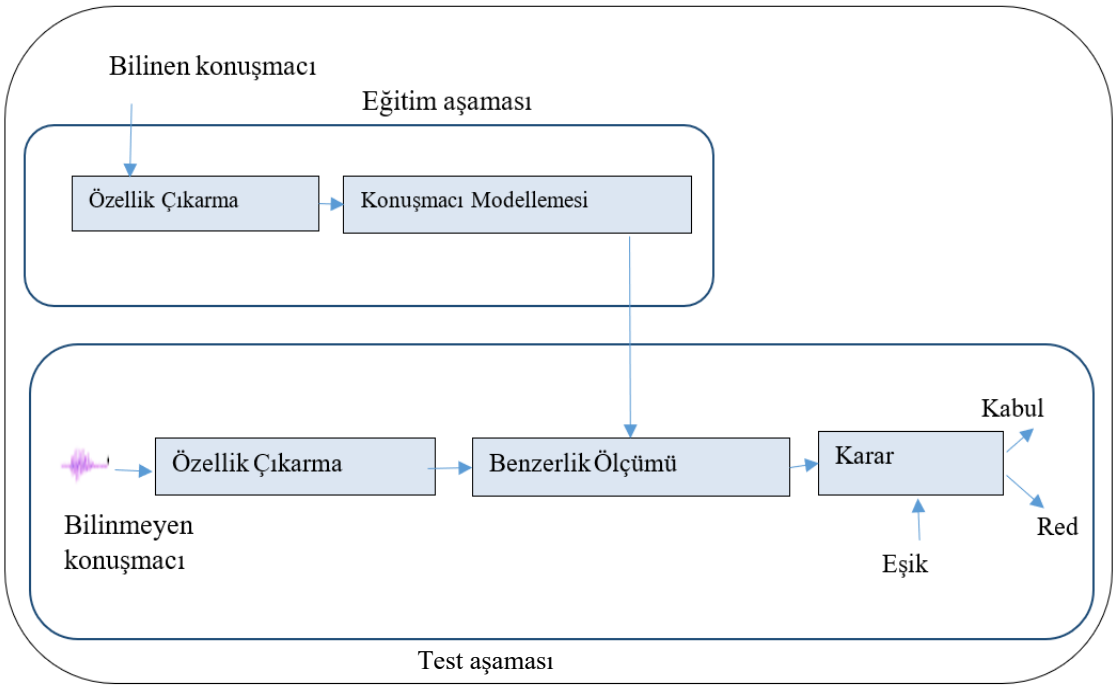
Konuşmacı doğrulama, konuşmacının kimliğini doğrulamak için ses analizi temelinde yapılan bir süreçtir. Bu süreç, bir kişinin önceden kaydedilmiş sesiyle gerçek zamanlı olarak verilen sesi karşılaştırarak kimliği doğrular.

Kullanılan sınıflama teknikleri arasında:

- **Gaussian Mixture Model (GMM):** Ses özelliklerini bir dizi Gauss dağılımı kullanarak temsil eder. Eğitim verilerine dayalı olarak, konuşmacının modelini oluşturarak yeni sesleri sınıflandırır.
- **Support Vector Machine (SVM):** Ses sinyallerini sınıflandırmak için kullanılan başka bir popüler yöntemdir. Ses özelliklerini kullanarak sınıflandırma modeli oluşturur ve yeni sesleri tanımlamak için bu modeli kullanır.
- **Deep Neural Networks (DNN):** Yapay sinir ağları, konuşmacı doğrulamada başarıyla kullanılan güçlü bir tekniktir. Ses sinyallerini temsil etmek ve sınıflandırmak için derin öğrenme modelleri kullanır.
- **Convolutional Neural Networks (CNN):** Özellikle konuşma sinyalleri için spektrogramlar gibi görsel temsil yöntemlerini işlemek için kullanılır.

Bu teknikler, konuşmacı doğrulama sistemlerinin başarısını artırmak için bir arada kullanılabilir veya tek başına kullanılabilir. Ancak her bir yöntemin avantajları ve dezavantajları vardır ve uygulama senaryosuna göre tercih edilebilir.

Konuşmacı doğrulama yöntemleri metne bağımlı ve metinden bağımsız yöntemlere ayrılır. Metne bağlı yöntemlerde, konuşmacı doğrulama sistemi konuşulacak metin hakkında önceden bilgiye sahiptir ve kullanıcının bu metni konuşması beklenir. Ancak metinden bağımsız bir sistemde, sistemin konuşulacak metin hakkında önceden bilgisi yoktur ve kullanıcının metni konuşması beklenmez. Metne bağımlı sistemler nispeten kısa konuşmalardan yüksek konuşmacı doğrulama performansı elde ederken, metinden bağımsız sistemler güvenilir modelleri eğitmek ve iyi performans elde etmek için uzun konuşmalar gerektirir.



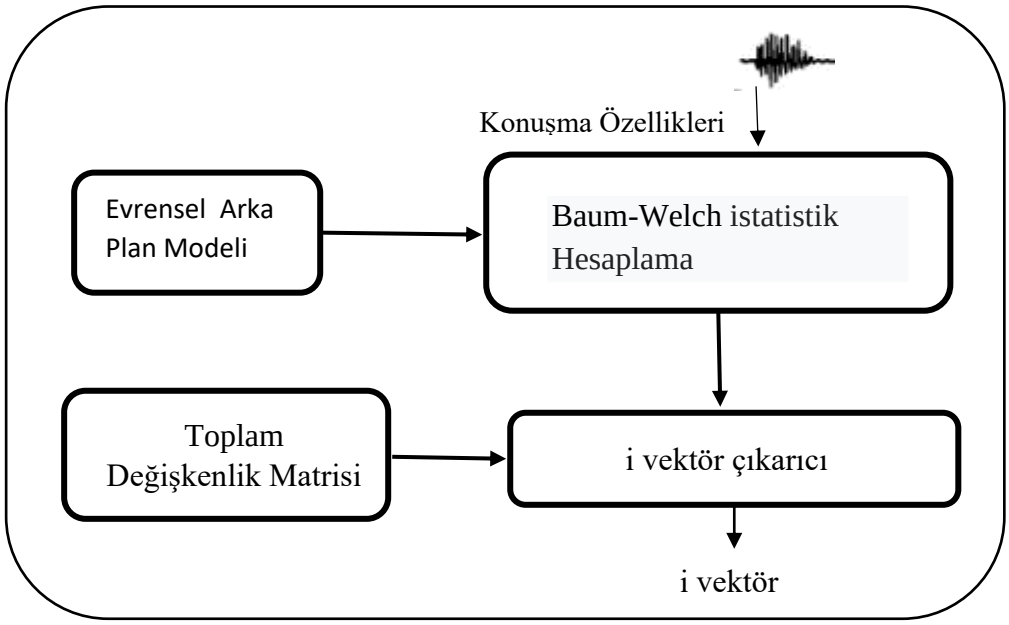

**Şekil 3.1.** Temel bir konuşmacı doğrulama sisteminin blok diyagramı

Temel bir konuşmacı doğrulama sisteminin yukarıdaki blok diyagramında gösterildiği gibi, bir konuşmacı doğrulama sistemi iki ana aşamadan oluşur: hedef konuşmacıların kaydedildiği *eğitim aşaması* ve konuşmacının kimliği hakkında bir kararın alındığı *test aşaması*. Eğitim açısından bakıldığında, konuşmacı modelleri üretici ve ayrımcı olarak sınıflandırılabilir. Gauss Karışım Modeli (GMM) gibi üretken modeller, her konuşmacıdaki özellik dağılımını tahmin eder. Destek Vektör Makinesi ve Derin Sinir Ağı (Deep Neural Net (DNN)) gibi ayrımcı modeller, aksine, konuşmacılar arasındaki sınırı modeller.

Konuşmacı doğrulama sistemlerinin performansı, kayıt ve doğrulama konuşma sinyalleri arasındaki kanallardaki ve oturumlardaki değişkenlik nedeniyle düşer. Kanal/oturum değişkenliğini etkileyen faktörler şunlardır:

1. Kayıt ve doğrulama konuşma sinyalleri arasındaki kanal uyumsuzluğu, örneğin kayıt ve doğrulama konuşma sinyallerinde farklı mikrofonlar kullanma.
2. Çevresel gürültü ve yankılanma koşulları.
3. Yaşlanma, sağlık, konuşma tarzı ve duygusal durum gibi konuşmacı sesindeki farklılıklar.
4. Sabit hat, cep telefonu, mikrofon ve İnternet protokolü (VoIP) üzerinden ses gibi iletim kanalı.

Konuşmacı doğrulama ve belirleme için günümüzde en yaygın olarak kullanılan metotlardan birisi i-vektör metodudur. GMM-UBM'ye dayanan i-vektör yaklaşımı en popüler yaklaşımlardandır. i-vektörlerin en yaygın kullanılan özellik normalleştirme tekniklerinden biri uzunluk normalleştirmedir. Uzunluk normalleştirme, i-vektörlerin dağılımının Gauss normal dağılımıyla eşleşmesini sağlar ve i-vektörünün dağılımlarını daha benzer hale getirir. Uzunluk normalleştirmeden önce beyazlatma işlemi yapmak, konuşmacı doğrulama sistemlerinin performansını artırır. i-vektör normalizasyonu, i-vektörlerin gaussiyanitesini geliştirir ve verilerin altında yatan varsayımlar ile gerçek dağılımlar arasındaki boşluğu azaltır. Ayrıca, değerlendirme ve test i-vektörleri arasındaki veri kümesi değişimini de azaltır.




**Şekil 3.2.** i -vektör çıkarma işlemi

i-vektörlerinin elde edilmesinde en yaygın iki teknik, Sınıf İçi Kovaryans Normalizasyonu (WCCN) ve Doğrusal Diskriminant Analizi'dir (LDA).

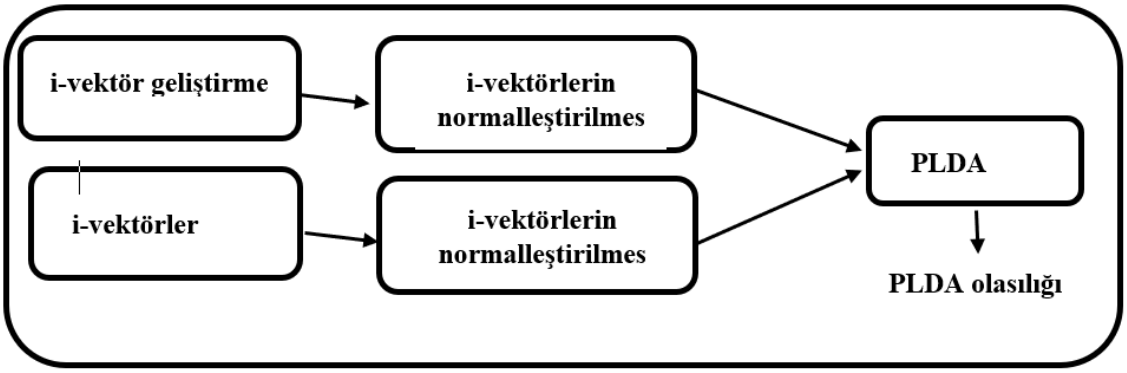
WCCN değişkenliği telafi etmek amacıyla kosinüs çekirdeği işlevlerini normalleştirmek için sınıf içi kovaryans matrisini kullanır. LDA ise kanal efektlerinin neden olduğu konuşmacı içi değişkenliği en aza indiren ve konuşmacı arasındaki değişkenliği en üst düzeye çıkaran azaltılmış özel eksenler tanımlamaya çalışır.

Konuşma kümelerinin çıktılarından i-vektörler çıkarıldıktan sonra, kosinüs mesafesi puanlaması, iki i-vektörün aynı konuşmacıya veya farklı konuşmacılara ait olup olmadığı hipotezini test eder. İki i-vektör verildiğinde, aralarındaki kosinüs mesafesi aşağıdaki gibi hesaplanır (George, K. K ve ark., 2015)

$$cos\left(\omega_i, \omega_j\right) = \frac{\omega_i \omega_j}{\|\omega_i\| \cdot \|\omega_j\|} \gtrless \theta \tag{3.1}$$

burada θ eşik değeridir ve $cos\left(\omega_i, \omega_j\right)$, i ve j kümeleri arasındaki kosinüs mesafesi puanıdır. İ ve j kümeleri için çıkarılan karşılık gelen i-vektörleri sırasıyla $\omega_i$ ve $\omega_j$ ile temsil edilir. Kosinüs mesafesi puanlaması, büyüklüklerini değil, yalnızca iki i-vektör arasındaki açıyı dikkate alır. Oturum ve kanal değişkenlikleri gibi konuşmacı olmayan bilgiler i-vektör büyüklüğünü etkilediğinden, büyüklüklerin kaldırılması i-vektör sistemlerinin sağlamlığını artırır.

Olasılıksal doğrusal diskriminant analizi (PLDA) modelleme tekniği tarafından takip edilen i-vektör gösterimi, konuşmacı doğrulama sistemlerinde en son teknolojidir. PLDA, konuşmacı tanıma deneylerinde başarıyla uygulanmıştır. Ayrıca, konuşmacı doğrulama görevinde konuşmacı ve oturum değişkenliğini ele almak için de uygulanır.




**Şekil 3.3**. PLDA modeli örneği

PLDA modeli Gauss davranışını varsaymasına rağmen, kanal ve konuşmacı etkilerinin Gaussian olmayan i-vektörlerle sonuçlandığına dair ampirik kanıtlar vardır. Öğrencinin t-dağılımının, varsayılan Gauss PLDA modelinde kullanılmasının performansı artırdığı bildirilmiştir.

Bu normalleştirme tekniği karmaşık olduğundan, radyal Gaussianizasyon adı verilen i-vektörlerinin doğrusal olmayan bir dönüşümü önerilmiştir. i-vektörlerini beyazlatır ve uzunluk normalleştirmesi gerçekleştirir. Bu, PLDA modelinin Gauss varsayımlarını geri yükler.

Gauss PLDA (GPLDA) adı verilen PLDA modelinin bir varyantının daha iyi sonuçlar sağladığı gösterilmiştir. Düşük hesaplama gereksinimleri ve performansı nedeniyle, en yaygın kullanılan PLDA modellemesidir. GPLDA modelinde, konuşmacı içi değişkenlik, kanal alt alanını atlamamıza izin veren tam bir kovaryans kalıntısı terimi ile modellenir.

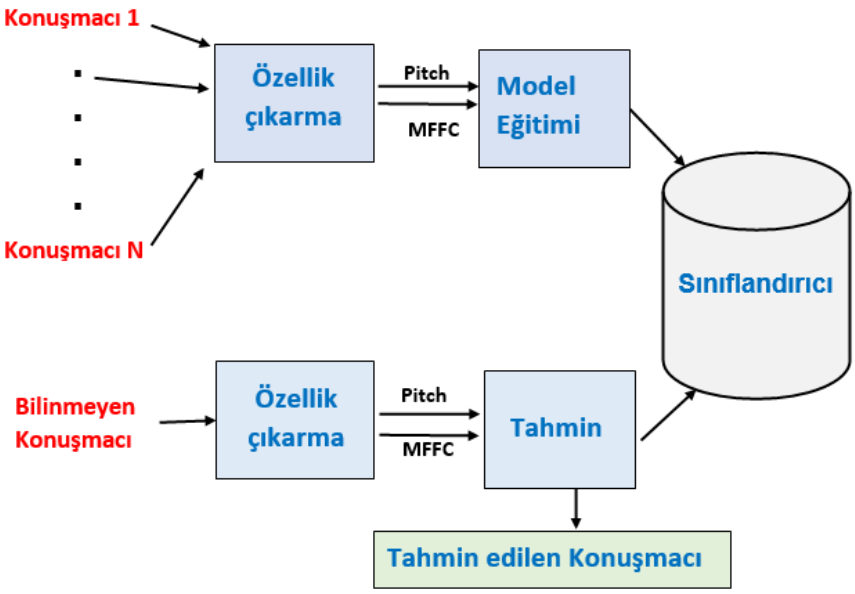
### 3.2. Konuşmacı Belirleme

Konuşmacı belirleme, doğal dil işleme alanında önemli bir konudur ve bir metinde belirli bölümlerin hangi konuşmacılara ait olduğunu tanımlamayı amaçlar. Özellikle metin tabanlı diyalog sistemlerinde ve sosyal medya analizlerinde kullanılır. Konuşmacı belirleme için kullanılan sınıflama teknikleri arasında şunlar bulunabilir:

- **Kural Tabanlı Yöntemler:** Metinde belirli kural ve özellikler kullanarak konuşmacıyı tespit etmeye çalışan basit yaklaşımlardır.
- **Makine Öğrenmesi Yöntemleri:** Metin sınıflandırma problemleri için sıkça kullanılan yöntemler arasında yer alır. İlgili özniteliklerin belirlenmesi ve makine öğrenmesi algoritmalarının kullanılmasıyla konuşmacı belirlemeyi gerçekleştirir.
- **Doğal Dil İşleme Yöntemleri:** Metindeki dilbilgisi yapıları ve anlam ilişkilerini dikkate alan yöntemlerdir. Çekimli fiil kullanımı, zamirler ve diğer dilbilgisi özellikleri konuşmacı belirlemede önemli rol oynayabilir.
- **Özellik Tabanlı Yöntemler:** Metindeki farklı öznitelikleri (örneğin, kelime sıklığı, kelime uzaklığı, cümle yapıları vb.) kullanarak konuşmacı belirlemeyi gerçekleştirir.
- **Derin Öğrenme Yöntemleri:** Son zamanlarda popüler hale gelen ve metin sınıflandırma alanında oldukça etkili olan derin öğrenme modelleri, konuşmacı belirlemede de kullanılabilir. Özellikle LSTM (Uzun-Kısa Süreli Bellek) ve GRU (Geri Beslemeli Birim) gibi tekrarlanan sinir ağları kullanılarak başarılı sonuçlar elde edilebilir.

Bu tekniklerin kombinasyonu veya farklı metin yapılarına uygun özelleştirilmiş yöntemler, konuşmacı belirleme problemlerinde başarılı sonuçlar elde etmek için kullanılabilir.

Aşağıdaki Şekil 3.4'te konuşmacı belirleme için kullanılan yaklaşım diyagramda gösterilmiştir. Bu şekil kaydedilen konuşmadan elde edilen özelliklere dayanarak kişileri belirlemek için bir sınıflama yaklaşımını göstermektedir. Sınıflandırıcıyı eğitmek için kullanılan özellikler, konuşmanın seslendirilen bölümlerinin perdesi ve mel frekans spektrum katsayılarıdır (MFCC). Bu kapalı sistem bir konuşmacı belirlemesidir. Test edilen konuşmacı sesi, mevcut tüm konuşmacı modelleriyle (sonlu bir küme) karşılaştırılır ve en yakın eşleşmeye ait konuşmacı seçilir.




**Şekil 3.4.** Konuşmacı belirleme için kullanılan yaklaşım diyagramı

# 4. MATERYAL

Konuşmacı belirleme çalışmalarında METUBET ve MNIST veri tabanı kullanılmıştır. METUBET veri tabanı Türkçe dilindeki tüm fonemleri içermektedir. Yaklaşık 500 dakikalık ses verisi mevcuttur. Her konuşmacı yaklaşık 40 cümle okumuştur ve 2462 özgün cümleden oluşmaktadır. Veriler 68'i erkek 52'si kadın 120 kişinin ses kayıtlarını içerir. Ses işareti Hamming pencere uygulanarak 15 ms'lik kısmi örtüşen 30 ms uzunluğundaki çerçevelere ayrılarak işlenmiştir. Ses işareti pencerelendikten sonra Hızlı Fourier Dönüşümü (HFD) alınarak, elde edilen vektör mel ölçekte 0-8000 Hz arasına yerleştirilmiş ve üçgen süzgeç takımına uygulanmıştır. Perde (Pitch) ve MFCC katsayıları konuşmacıları sınıflandırmak için kullanılan iki özelliktir. Sıfır geçiş hızı ve kısa süreli enerji ise, perde özelliğinin ne zaman kullanılacağını belirlemek için kullanılır. Her bir çerçeveye karşılık olarak METUBET ve MNIST'de 14, 40 boyutlu öznitelik vektörleri elde edilmiştir. Bu öznitelikler MFCC ve onun türevleri alınarak oluşturulmuştur. Konuşmacı tanıma deneylerinde METUBET için 20 erkek, 20 bayan 40 konuşmacı, MNIST için ise 15 erkek 15 bayan olarak 30 konuşmacı kullanılmıştır. Konuşmacı belirlemede sınıflayıcı olarak zaman gecikmeli sinir ağı (TDNN), KNN ve SVM ve OVY'den faydalanılmıştır. Deneysel çalışmalar MATLAB programlama dili ile gerçekleştirilmiştir.

## 4.1. Sinyal Sınırlarının Bulunması

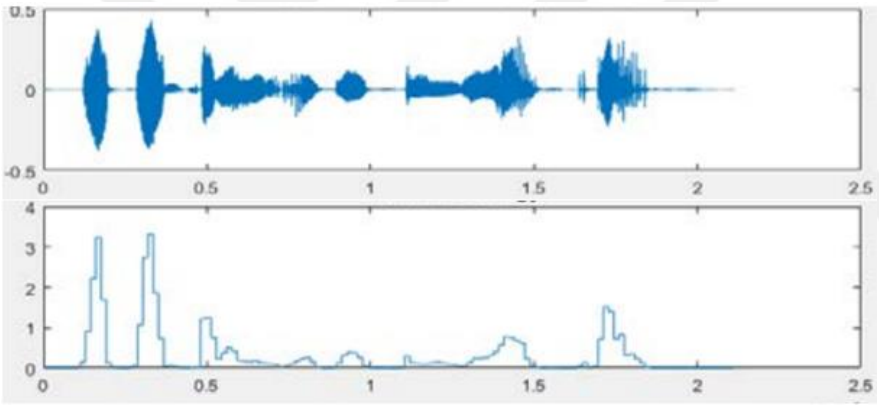
Öz vektörler elde edilmeden önce ses sinyallerine ön işlem uygulamak gerekir. Bu adımlar aşağıda verilmiştir. Temel prensip olarak konuşma tanımada ses sinyalinin gereksiz kısımlarının alınmayıp sadece kelime sinyalinin alınması gerekir. Bunun için ses sinyalinin yalnızca gürültülü bölümlerinden belirlenen eşik değeri ile karşılaştırılmasıyla sesli (kelime) bölümü ve gürültülü bölümü tespit edilebilir. Bunun için sıfır geçiş oranı (SGO) ve kısa süreli enerji (KSE) değerleri birlikte kullanılarak başarılı biçimde sesin sınırlarının tespiti gerçekleştirilebilmektedir.

### 4.1.1. Kısa Süreli Enerji (KSE)

Konuşma sinyallerinde zamana bağlı olarak konuşmanın genliği değişiklik göstermektedir. Konuşmanın olduğu bölgelerde KSE, konuşmanın olmadığı bölgelere göre daha yüksektir. Bu sayede belli bir enerji eşik değeri kullanarak kelime sinyalinin sınırları belirlenebilir. Genel olarak 30 ms gibi küçük pencere süreleri kullanılarak hesaplanırlar.

### 4.1.2 Sıfır Geçiş Oranı

Sıfır geçiş oranı, sinyalin bir çerçevesi boyunca değiştirdiği işaret oranıdır. Diğer bir deyişle, bir çerçeve boyunca sinyalin genlik değerlerinin kaç kez pozitiften negatife (ya da negatiften pozitife) değiştiğiyle alakalıdır. Bu değişim sayısı çerçeve uzunluğuna bölünerek sıfır geçiş oranı bulunur (Çolak, R., ve Akdeniz, R. 2018).




**Şekil 4.1.** Sinyalin sıfır geçiş grafiği

## 4.2. Öznitelik Çıkarma

Ses tanıma sistemlerinin ilk aşamasında, ses sinyali konuşmacı özelliklerini temsil eden daha az değişkenlik taşıyan ve daha fazla ayırıcı özellik içeren parametrik değerlere dönüştürülür. Bu parametrik değerler, ses sinyalinden çıkarılmasında kullanılan çeşitli yöntemlerle öznitelik olarak adlandırılır. Bu yöntemler, kısa süreli analiz yöntemleri olarak bilinir ve ses sinyalinin durağan olarak kabul edilen kısa parçaları üzerinde uygulanır. Bu süreç sonucunda, her analiz parçasından elde edilen değerler birleştirilerek öznitelik vektörü oluşturulur. Bu çalışmada, öznitelik vektörü olarak Mel Frekans Kepstrum Katsayıları (Mel Frequency Cepstrum Coefficient (MFCC)) kullanılmıştır.

### 4.2.1. Ön Vurgulama

Yüksek frekanslı bölgelerde ses sinyalinin genliğinde bir sönümleme oluşur. Bu nedenle ön vurgulama olarak isimlendirilen ve yüksek frekanslı bölgeleri güçlendiren bir filtreleme uygulanır. Denklem 1 ifadesiyle temsil edilen transfer fonksiyonuna sahip FIR filtresi kullanılır (Keser, S., ve Edizkan, R. 2009).

$$H(z) = 1 - \alpha z^{-1} \quad 0.9 \leq \alpha \leq 1 \tag{4.1}$$

### 4.2.2. Çerçeveleme ve Pencereleme

Yaklaşık durağan özelliklere sahip olduğu görüldüğünde, genellikle konuşma sinyalleri kısa süreli parçalara (20-30 ms) ayrılarak işleme tabi tutulur. Bu parçalar, çerçeve adı verilen birimlerdir ve çerçevelerin sınırlarındaki bilgi kaybını engellemek için önceki çerçevenin arka kısmıyla (10-20 ms) örtüşür. Sinyalin kenarlarında meydana gelen kesinti, çerçevelere pencere fonksiyonu uygulanarak azaltılır. Bu amaçla genellikle Hamming penceresi tercih edilir (Han, W. et al., 2006).

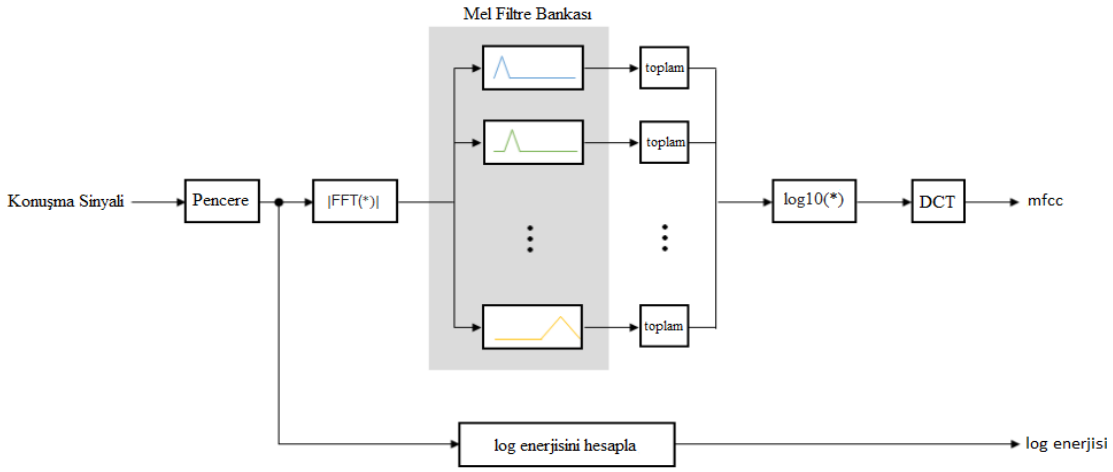
$$w[n] = (1 - \alpha) - \alpha cos\ (\frac{2\pi n}{L-1}), \tag{4.2}$$

Burada L pencere uzunluğu ve α = 0,46164 değerini alır.

### 4.2.3. Spektral Analiz ve MFCC Katsayılarının Bulunması

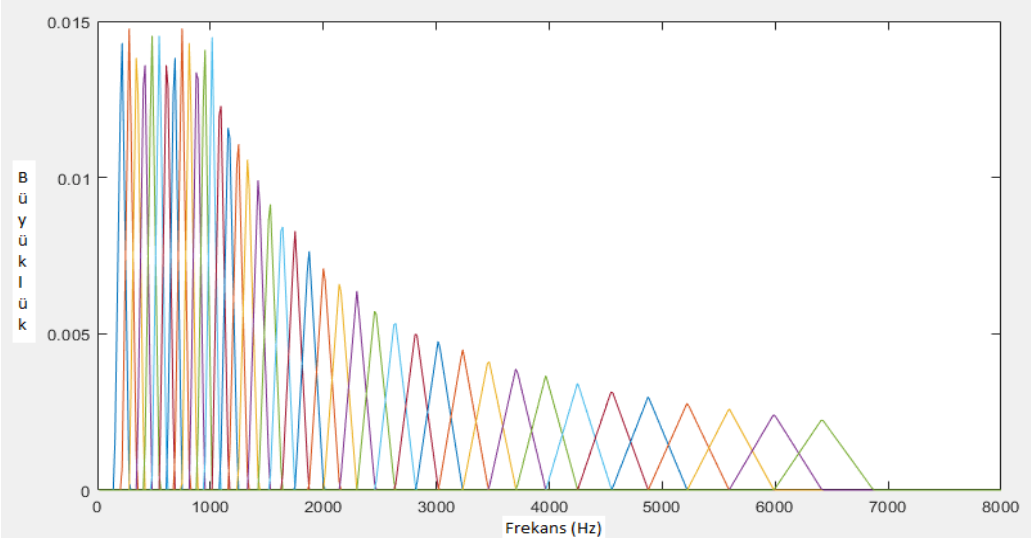
MFCC, ses işleme ve konuşma tanıma alanında sıkça kullanılan öznitelik çıkarma yöntemidir. Konuşma sinyallerini temsil etmek ve işlemek için kullanılır. MFCC, ses sinyallerini insan işitme özelliğine benzer bir şekilde temsil etmek amacıyla özellikle sesin frekans bileşenlerinin insan kulağının tepkisine uygun hale getirilmesini sağlar.

MFCC, tanıma görevlerinde kullanılmak üzere konuşma sinyallerinden çıkarılan popüler özelliklerdir. Konuşmanın kaynak-filtre modelinde, MFCC'nin filtreyi (ses yolu) temsil ettiği anlaşılmaktadır. Ses yolunun frekans tepkisi nispeten pürüzsüzdür, oysa sesli konuşmanın kaynağı bir dürtü treni olarak modellenebilir. Sonuç, ses yolunun bir konuşma segmentinin spektral zarfı ile tahmin edilebilmesidir. MFCC'nin motive edici fikri, kokleanın anlaşılmasına bağlı olarak ses yolu (düzleştirilmiş spektrum) hakkındaki bilgileri az sayıda katsayıya sıkıştırmaktır. MFCC'yi hesaplamak için katı bir standart olmamasına rağmen, temel adımlar diyagramda özetlenmiştir.




**Şekil 4.2.** MFCC katsayılarının elde edilme aşamaları

Mel filtre bankası ilk 10 üçgen filtreyi doğrusal olarak yerleştirir ve kalan filtreleri logaritmik olarak yerleştirir. Bireysel bantlar eşit enerji için ağırlıklandırılır. Grafik tipik bir mel filtre bankasını temsil eder.




**Şekil 4.3.** Kullanılan Mel filtre yapısı

MFCC, aşağıdaki adımları içeren bir süreçtir:

- **Ön vurgulama:** Giriş ses sinyaline ön-vurgulama filtresi uygulanarak yüksek frekans bileşenleri vurgulanır.
- **Çerçeveleme:** Ses sinyali küçük zaman çerçevelerine bölünür ve her bir çerçeve üzerinde işlemler yapılır.

- **Hamming Pencereleme:** Her bir zaman çerçevesi, sinyallerin hatalı analizini önlemek için Hamming penceresiyle çarpılır.
- **Fourier Dönüşümü:** Zaman çerçevelerine uygulanan FFT (Hızlı Fourier Dönüşümü) sayesinde frekans bileşenleri elde edilir. Bu değerler insan duyusunun iyi algıladığı 0-10 kHz frekans değer aralığı için bulunur.
- **Mel Filtre Bankası:** Ses sinyalinin frekans bileşenleri üzerinde Mel ölçeğine göre filtreleme yapılır. İnsan duyusu ses sinyalinin frekansını 1000 Hz'e kadar doğrusal, 1000 Hz'in üstündeki frekansları ise logaritmik olarak algılar. Bu yüzden, frekans içeriği doğrusal olmayan Mel ölçeği olarak isimlendirilen bir ölçek kullanılır. Yaklaşık 1000 Hz'ye kadar doğrusal; üzerinde ise logaritmik olarak değişen Mel ölçeği genellikle Denklem 3 ifadesiyle temsil edilir (Tiwari, V., 2010).

$$Mel(f) = 2595 \log10 (1 + 700\, f\,) \tag{4.3}$$

Bu ifadedeki $f$ gerçek frekansı $Mel(f)$ ise algılanan frekansı temsil etmektedir. Doğrusal olmayan bu frekans algılamasının MFCC özniteliklerine dahil edilmesi için sesin genlik spektrumu Mel ölçeğine göre eşit aralıklarla yerleştirilmiş bir filtre kümesiyle saptırılır. Bu filtre kümesi alt sınırı kendinden önceki, üst sınırı ise kendinden sonraki filtrenin merkez frekansına gelecek şekilde yerleştirilen üçgen filtrelerden oluşur. Ardından logaritmik sıkıştırma ile sesin gürlüğünün veya sessizliğinin filtre çıkışlarına etkisini azaltmak ve insanın doğrusal olmayan genlik hassasiyetini modellemek için filtre çıkışlarının logaritması alınır.

- **Cepstral Dönüşüm:** Mel filtreleri üzerinde logaritmik dönüşüm ve DCT (Discrete Cosine Transform) işlemi uygulanarak cepstral katsayıları elde edilir. Algoritmanın son aşamasında filtre çıkışlarına ayrık kosinüs dönüşümü uygulanarak yüksek korelasyona sahip katsayılar korelasyonsuz bir düzleme atanır. DCT sonucunda elde edilen vektörün sadece düşük dereceli katsayıları ile MFCC öznitelik vektörü oluşturulur.

Sonuç olarak, MFCC, ses sinyallerini insan işitme özelliğine uygun şekilde temsil eden bir dizi cepstral katsayısından oluşur. Bu katsayılar, konuşma tanıma ve diğer ses işleme uygulamalarında öznitelik olarak kullanılır.

### 4.2.4. Perde Frekansının Bulunması

Konuşma, sesli ve seslendirilmemiş şekillerde sınıflandırılabilir. Sesli konuşmada, akciğerden gelen hava ses telleri tarafından modüle edilir, bu da yarı-periyodik bir uyarılmaya yol açar. Bu uyarılmanın sonucu olarak, düşük frekansta salınan perde adı verilen bir ses oluşur. Öte yandan, seslendirilmemiş konuşmada akciğerden gelen hava daralmalarla geçer, bu da gürültüye benzer bir uyarılmaya dönüşür. Konuşmanın kaynak-filtre modelinde, kaynak sesi, filtre ise ses yolunu temsil eder. Kaynağın özelliklerini belirlemek, konuşma sistemini anlamak için önemlidir. Bir örnek olarak, "üç" kelimesini ele alalım. /ç/ ünsüzü (seslendirilmemiş) daha gürültülüyken, /ü/ ünlüsü (sesli) güçlü bir temel frekansta karakterize edilir (Eskidere, Ö., ve Ertaş, F., 2009).

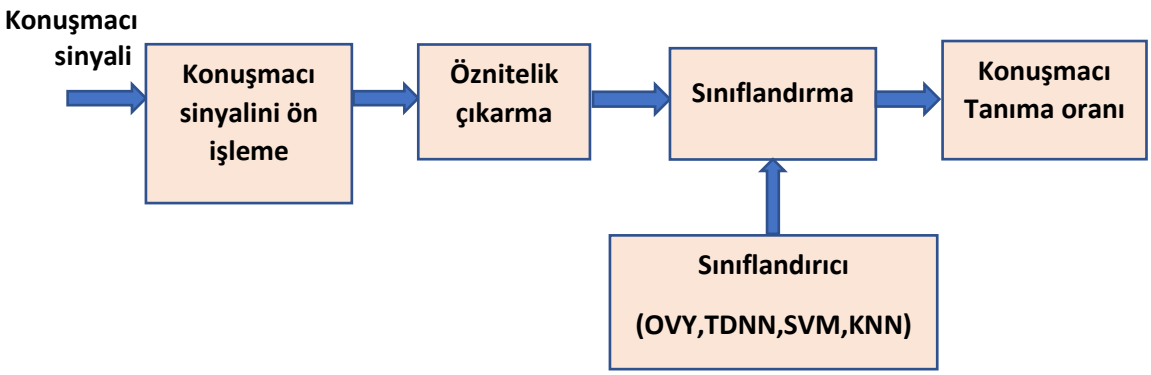
Konuşma sinyali doğal olarak dinamik ve zaman içinde değişir. Konuşma sinyallerinin kısa zaman dilimlerinde sabit olduğunu ve genellikle 20-40 ms'lik pencere boyutlarıyla işlendiğini kabul ederiz. Örneğin, bu çalışmada 30 ms'lik ardışık pencerelerin her biri 25 ms'lik örtüşmeli bir zaman aralığı kullanılmıştır. Perdenin zaman içindeki değişimini incelemek için perde işlevi kullanılır.

Perde işlevi, her pencere için bir perde değeri tahmin eder. Ancak perde sadece sesli konuşma bölgelerindeki bir kaynağın özelliğini yansıtır. Sessizlik ile konuşma arasındaki farkı ayırt etmenin en basit yolu, kısa zaman enerjisini incelemektir. Eğer bir penceredeki enerji belirli bir eşiğin üzerinde ise, o pencereyi konuşma olarak etiketlersiniz.

Sesli ve seslendirilmemiş konuşma arasındaki ayırt etmenin bir diğer temel yöntemi, sıfır geçiş oranını analiz etmektir. Çok fazla sıfır geçişi varsa, baskın düşük frekansta salınım olmadığını gösterir. Bir pencere için sıfır geçiş hızı belirli bir eşiğin altında ise, bu pencereyi sesli olarak tanımlarsınız.

## 4.3. Konuşma Tanımada Kullanılan Sınıflayıcılar

Eğitim aşamasında her konuşma sinyali sınıfına ait öznitelikler kaydedildikten sonra test aşamasına geçilir. Test aşamasında test sinyaline ait öznitelik vektörü ile eğitim kümesindeki her sınıfa ait öznitelikler bir sınıflandırıcı ile karşılaştırılarak test sinyali bir sınıfa atanır. Burada her sınıf bir konuşmacıya denk gelmektedir. Çalışmada test aşamasında kullanılan genel konuşmacı tanıma sistemi aşağıdaki Şekil 4.4'de verilmiştir.




**Şekil 4.4.** Test aşamasında kullanılan genel konuşmacı tanıma sistemi

### 4.3.1. Destek Vektör Makineleri (Support Vector Machine (SVM))

Destek vektör makineleri (SVM), güçlü bir sınıflandırıcı olarak iki sınıf arasında ayırıcı bir hiper düzlem bulma amacı taşır. Bir veri kümesinin doğrusal olarak sınıflandırılabildiği durumlarda, ayırıcı hiper düzlem sayısının birden fazla olabileceği belirtilmiştir (McLaren, M. ve ark.., 2010). SVM'nin temel amacı, maksimum marjine sahip bir hiper düzlem tespit etmektir. SVM sınıflandırıcısı, doğrusal olmayan sınıflandırma problemlerini çözebilmek amacıyla çekirdek fonksiyonları kullanabilir. Çekirdek fonksiyonları, giriş vektörlerini yüksek boyutlu bir uzaya taşımak amacıyla kullanılır. Bu sayede, giriş uzayında doğrusal olarak sınıflandırılamayan veriler, genişletilmiş uzayda doğrusal olarak sınıflandırılabilir hale gelir. Lineer, RBF ve polinom çekirdekleri gibi farklı alternatifler, çekirdek fonksiyonu olarak tercih edilebilir.

### 4.3.2. K-en yakın komşu (k-*nearest neighbors,* (KNN))

En yakın komşu sınıflandırıcısı KNN için hiper parametreler arasında en yakın komşuların sayısı, komşulara olan mesafeyi hesaplamak için kullanılan mesafe metriği ve mesafe metriğinin ağırlığı bulunur (Umar, R. ve ark., 2019). Hiper parametreler, test kümesindeki doğrulama doğruluğunu ve performansını optimize etmek için seçilir. Bu algoritma kapsamında tahminde bulunmak istediğimiz gözlem birimine en yakın K adet farklı gözlem birimi tespit edilir ve bu K adet gözlem biriminin bağımlı değişkenleri üzerinden ilgili gözlem için tahminde bulunulur.

### 4.3.3. Zaman Gecikmeli Sinir Ağı (Structure of Time Delay Neural Network (TDNN) )

TDNN çok katmanlı yapay sinir ağı mimarisinin amacı kayma değişmezliği ile örüntüleri sınıflandırmak ve ağın her katmanında bağlamı modellemekdir (Liu, T., ve ark., 2022). Kayma değişmezliği sınıflandırmada, sınıflandırıcının sınıflandırmadan önce ses sinyalini bölümlemesine gerek yoktur. Konuşma gibi zamansal bir modelin sınıflandırılması için, TDNN böylece konuşma sinyalini sınıflandırmadan önce başlangıç ve bitiş noktalarını belirlemek zorunda kalmaz. Bir TDNN'de bağlamsal modelleme için, her katmandaki her bir sinir birimi, yalnızca bir katmandaki aktivasyonlardan/özelliklerden değil, aynı zamanda birim çıktı modelinden de girdi alır.

### 4.3.4. Ortak Vektör Yaklaşımı (OVY)

Ortak vektör yaklaşımı (OVY), ses tanıma ve görüntü tanıma alanlarında kullanılan bir alt uzay sınıflama yöntemini ifade eder. Bu metodun temelinde, her sınıfa ait değişmez özellikleri taşıyan bir vektör elde edilir ve bu vektör "ortak vektör" olarak anılır (Keser, S., ve Edizkan, R., 2009). Söz konusu durumda, eğitim setinde her biri k örnek içeren toplam c farklı sınıf olduğunda, eğitim kümesinin toplam örnek sayısı m=kc olacaktır. Burada m, ses komutu sınıfına ait vektör sayısını temsil ederken, n her bir vektörün boyutunu gösterir. OVY, yeterli veri durumlarında (m≥n) olduğu kadar, yetersiz veri durumlarında (m<n) da uygulanabilir. Aynı prensip, Adaptif Ortak Vektör Yaklaşımı (AOVY) için de geçerlidir. Sınıfı i olan r'inci sinyal örneği n-boyutlu uzayda $x_j^i$ olarak ifade edilirse, sınıflar arası dağılım matrisi $S_w$ aşağıdaki gibi hesaplanır:

$$S_w = \sum_{i=1}^{c} \sum_{r=1}^{k} \left( \left(x_r^i - \mu_i\right)\left(x_r^i - \mu_i\right)^T \right) \tag{4.4}$$

Burada, $\mu_i$ i'nci sınıfa ait ortalama vektörü temsil eder. İki dik alt uzay olan Farklılık alt uzayı **V** ve farksızlık alt uzayı $\mathbf{V}^{\perp}$ elde edilir. Farksızlık alt uzayı $\mathbf{V}^{\perp}$, $S_w$ matrisinin sıfır öz değerlerine karşılık gelen öz vektörler tarafından genişletilir. **V** ve $\mathbf{V}^{\perp}$ uzaylarının iz düşüm matrisleri sırasıyla **P** ve $\overline{\boldsymbol{P}}$ olarak gösterilirse, eğitim setindeki örneklerin $\mathbf{V}^{\perp}$ uzayındaki iz düşümleri aşağıdaki gibi ifade edilir.

$$\boldsymbol{x_{com}^i} = \boldsymbol{x_r^i} - \boldsymbol{P x_r^i} = \overline{\boldsymbol{P}} \boldsymbol{x_r^i}, \text{ r=1,2,…,N, i=1,2,…,c} \tag{4.5}$$

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

### 4.4. Sınıflandırma için Kullanılan Özellikler

Bu bölümde perde, sıfır geçiş hızı, kısa süreli enerji ve MFCC ele alınmaktadır. Perde (Pitch) ve MFCC, konuşmacıları sınıflandırmak için kullanılan iki özelliktir. Sıfır geçiş hızı ve kısa süreli enerji, perde özelliğinin ne zaman kullanılacağını belirlemek için kullanılır.

Pitch ve MFCC, N adetlik konuşmacılar için kaydedilen konuşma sinyallerinden çıkarılır. Pitch ve MFCC aynı ölçekte değildir. Bu, sınıflandırıcıyı sapmalı hale getirecektir. Ortalama değeri çıkararak ve standart sapmaya bölerek özellikler normalize edilir.

```
M = mean(features,1);

S = std(features,[],1);

features = (features-M)./S;
```

Elde edilen bu özellikler, sınıflayıcıları eğitmek için kullanılır. Ardından, sınıflandırılması gereken yeni konuşma sinyalleri aynı özellik ayıklamasından geçer. Sınıflayıcı, N adetlik konuşmacıdan hangisi en yakın eşleşme veriyorsa o konuşmacıyı seçer. Çalışmada 1 adet perde frekansı değeri ve 13 MFCC toplam 14 katsayı ve 1 adet perde frekansı değeri ve 39 MFCC toplam 40 katsayı olmak üzere 2 farklı katsayı vektörü kullanılmıştır.

# 5. BULGULAR VE TARTIŞMA

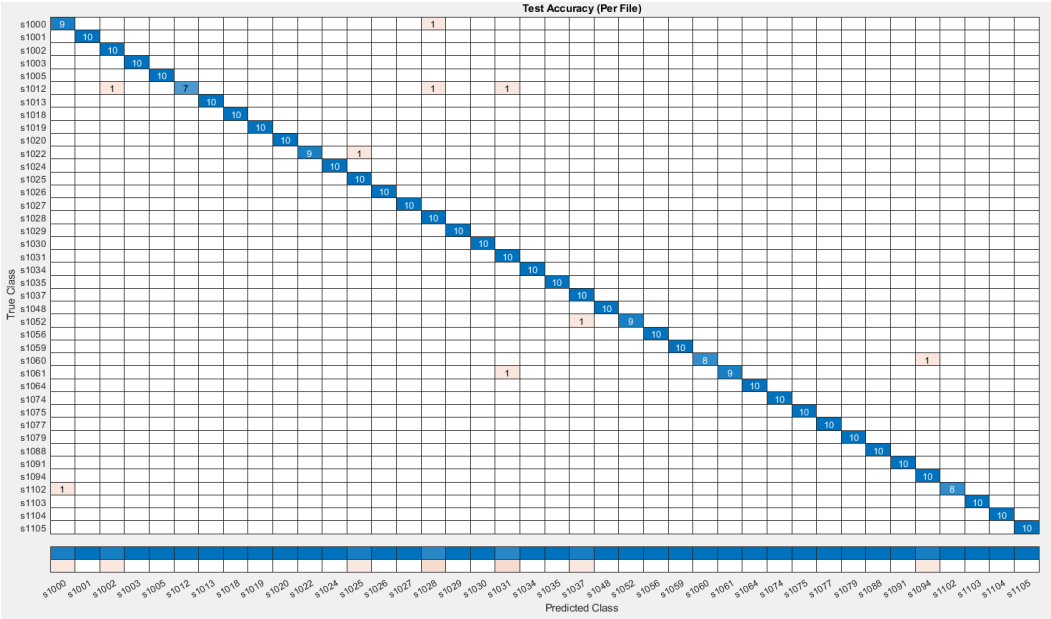
## 5.1 Deneysel Çalışmalar

Tüm ses sinyalleri eğitim ve test olarak ikiye ayrılmıştır. Çalışmamızda konuşmaların yalnızca ses içeren bölümlerinden elde edilen öznitelikler kullanılmıştır. Bu amaçla 30 ms genişliğindeki bir Hamming penceresinin 15 ms kaydırılmasıyla çerçevelere bölünen ses sinyalinden konuşma içermeyen kısımlar enerjiye dayalı olarak çıkarılmıştır. Daha sonra her çerçeveden elde edilen katsayılardan öznitelik vektörü oluşturulur. Çalışmamızda her bir çerçeveden elde edilen 13 MFCC katsayısı ve onların birinci ve ikinci türevlerinden oluşan toplam 39 katsayı öznitelik olarak kullanılmıştır. Ayrıca her çerçeveye ait bir perde frekansı değeri bulunarak MFCC katsayılarına eklenmiştir. Yani toplamda 14 ve 40 uzunluklu öznitelik vektörleri oluşturulmuştur. Sese sinyallerinden elde edilen MFCC ve perde değerleri her bir çerçeve için bulunarak alt alta sıralanmış ve ilgili konuşmacıya ait etiketle belirtilmiştir. SVM için bu çalışmada Lineer, RBF ve polinom çekirdekleri tercih edilmiştir. KNN için komşu sayısı 3 ve 5 olarak alınmış ve seçilen mesafe metriği Öklid uzaklık olarak seçilmiştir. TDNN için ise Tablo 5.1'de çalışmada uygulanan TDNN ağın mimarisini özetlenmektedir. Burada T, bir ses sinyalindeki toplam çerçeve sayısı ve N, eğitim setindeki sınıfların (konuşmacıların) sayısıdır.

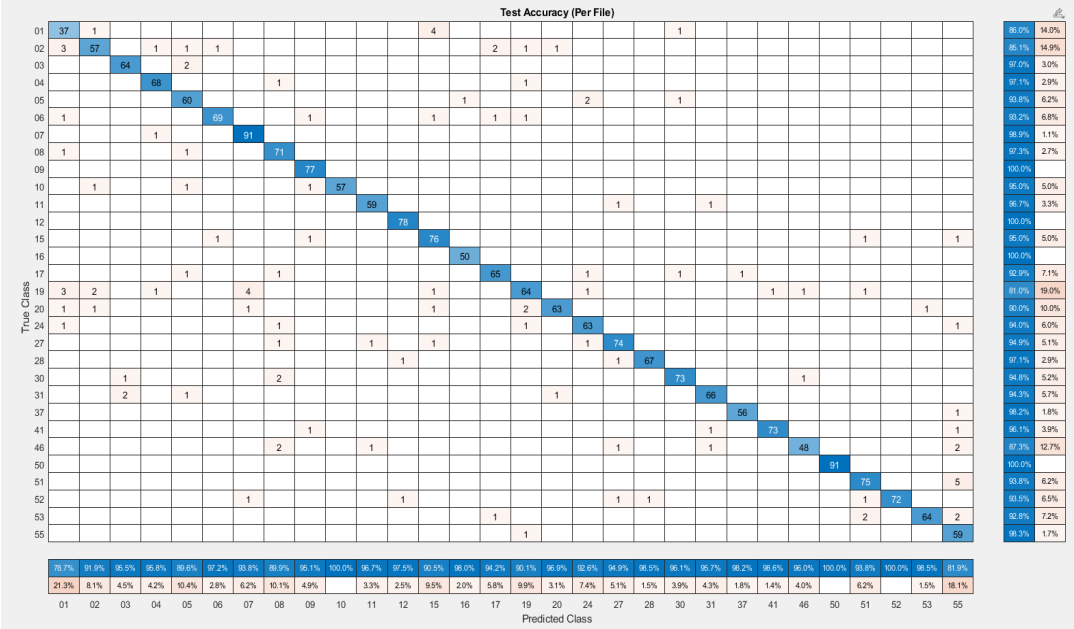
**Tablo 5.1.** Çalışmada uygulanan TDNN ağın mimarisi

| Katman | Açıklama | Katman içeriği | Toplam içerik | Giriş-Çıkış boyutu |
|---|---|---|---|---|
| 1 | 1-boyutlu konvolüsyon (Batch narmalizasyon+ReLu) | [t-2,t+2] | 5 | (5×Filtre sayısı) × Filtre sayısı |
| 2 | 1-boyutlu konvolüsyon (Batch narmalizasyon+ReLu) | {t-2,t,t+2} | 9 | (3×Filtre sayısı) × Filtre sayısı |
| 3 | 1-boyutlu konvolüsyon (Batch narmalizasyon+ReLu) | {t-3,t,t+3} | 15 | (3×Filtre sayısı) × Filtre sayısı |
| 4 | 1-boyutlu konvolüsyon (Batch narmalizasyon+ReLu) | {t} | 15 | Filtre sayısı × Filtre sayısı |
| 5 | 1-boyutlu konvolüsyon (Batch narmalizasyon+ReLu) | {t} | 15 | Filtre sayısı ×1500 |
| 6 | Statik Havuzlama | [0,T} | T | (1500×T) × 3000 |
| 7 | Full-connected (Batch narmalizasyon+ReLu) | {0} | T | 3000×Filtre sayısı |
| 8 | Full-connected (Batch narmalizasyon+ReLu) | {0} | T | Filtre sayısı × Filtre sayısı |
| 9 | Full-connected (Batch narmalizasyon+ReLu) | {0} | T | Filtre sayısı ×N |

OVY için öznitelik vektörlerinin sayısı vektör boyutundan büyük olduğu için yeterli veri durumu oluşmuştur. Bu durumda fark ve farksızlık alt uzayları birbirinden tam olarak ayırt edilemediğinden çalışmada farksızlık uzayını geren öz vektörler enerjinin belli oranları için bulunmuştur. Enerjinin yaklaşık %20-25 civarında alınmasının en iyi sonuçları verdiği görülmüştür. METUBET ve MNIST veri tabanları için SVM ile test aşamasında her konuşmacıya ait 10 cümle ve her cümleye ait çerçeveler için bulunan tanıma sonuçları Şekil 5.1'de verilmiştir.



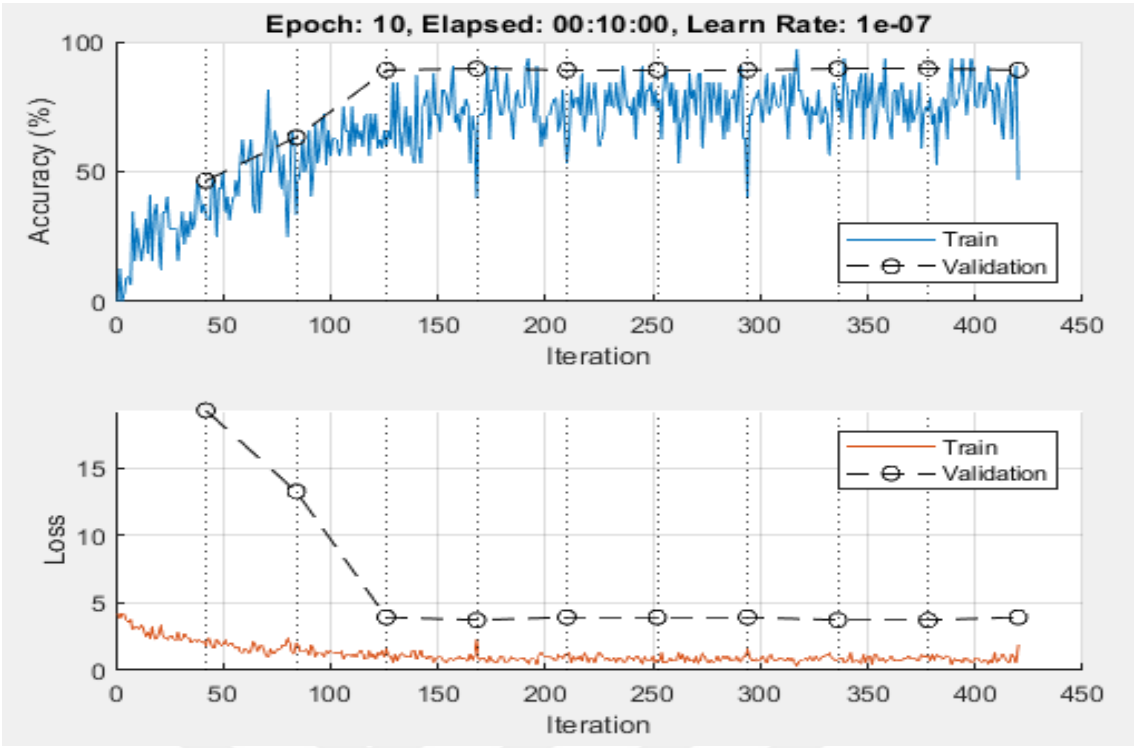

(a)



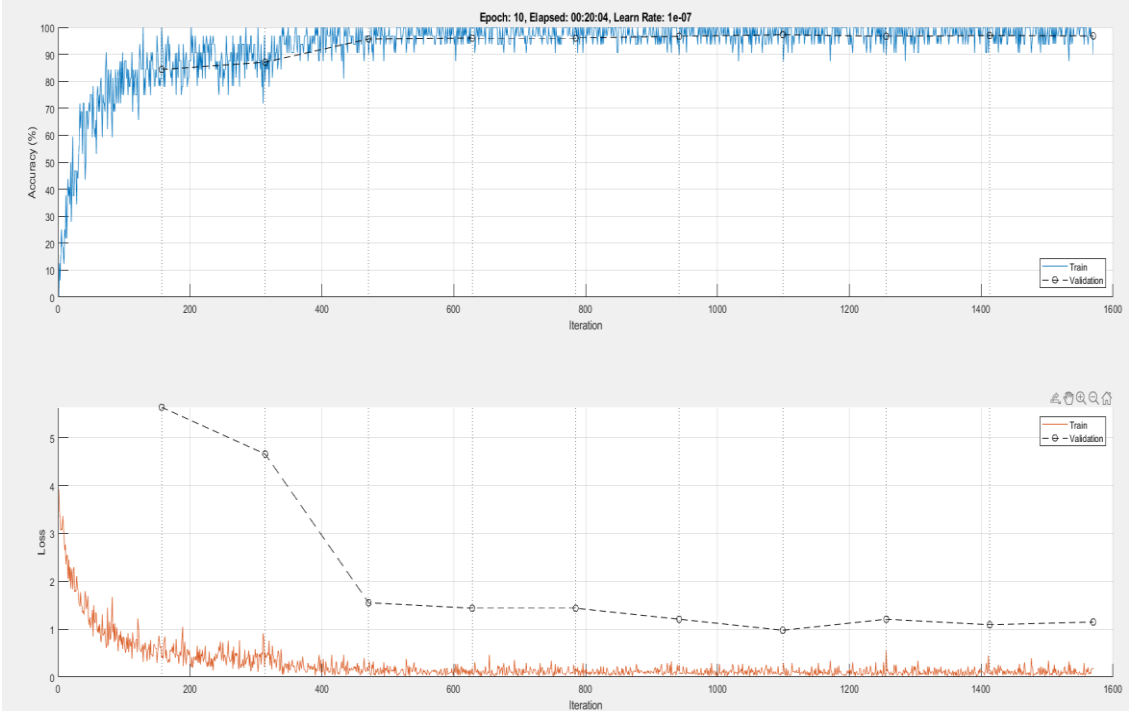

(b)

**Şekil 5.1** (a) METUBET ve (b) MNIST için SVM ile bulunan doğruluk oranları

METUBET ve MNIST veri tabanları için TDNN ile eğitim aşamasında elde edilen sonuç Şekil 5.2 (a) ve Şekil 5.2 (b) ‘de verilmiştir. METUBET’de test için her konuşmacının belli sürelerde seslendirdiği 10 cümle bulunmaktadır. Toplam 40 konuşmacı için 400 cümle sınıflandırılmıştır. MNIST veri tabanı için ise 30 kişi kullanılmıştır. Eğitim için her konuşmacıya ait 400 cümle, test için 100 cümle kullanılmıştır. Yani test için 3000 cümle bulunmaktadır. Konuşmacı belirlemede, METUBET veri tabanı için yapılan çalışma metinden bağımsız, MNIST veri tabanı için yapılan çalışma ise metne bağlı konuşmacı belirleme olarak kullanılmıştır.




(a)




(b)

**Şekil 5.2** (a) METUBET ve (b) MNIST için TDNN ile eğitim aşamasındaki entropi ve doğruluk oranları

Tablo 5.2 ve Tablo 5.3'te sırasıyla METUBET ve MNIST veri tabanları için farklı MFCC katsayı boyutlarıyla TDNN için bulunan doğruluk oranları verilmiştir.

**Tablo 5.2.** METUBET için TDNN ile test aşamasında elde edilen doğruluk oranları

| MFCC+pitch sayısı | Doğruluk oranı |
|---|---|
| 14 | 90.75 |
| 20 | 92.75 |
| 40 | 86.50 |

**Tablo 5.3.** MNIST için TDNN ile test aşamasında elde edilen doğruluk oranları

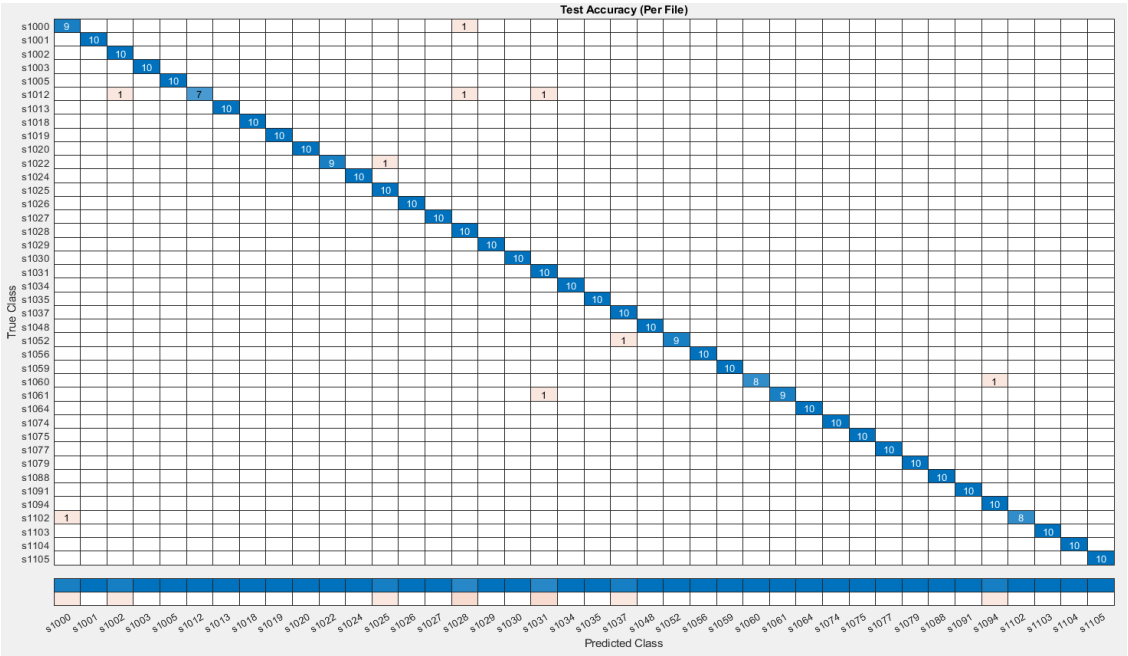
| MFCC+pitch sayısı | Doğruluk oranı |
|---|---|
| 14 | 90.04 |
| 20 | 96.17 |
| 40 | 85.58 |

METUBET veri tabanı için SVM ile 14 ve 40 boyutlu MFCC'ler için bulunan test doğruluk oranları Tablo 5.4'te verilmiştir.

**Tablo 5.4.** METUBET için SVM ile test aşamasında elde edilen doğruluk oranları

| MFCC+pitch sayısı | Polinom kernel | RBF kernel | Lineer kernel |
|---|---|---|---|
| 14 | 95.25 | 82.5 | 82.5 |
| 40 | 97.75 | 85.5 | 83.5 |

METUBET veri tabanı için SVM-Polinom kernel ile 40 boyutlu MFCC için bulunan test sonuçları Şekil 5.3'te verilmiştir.



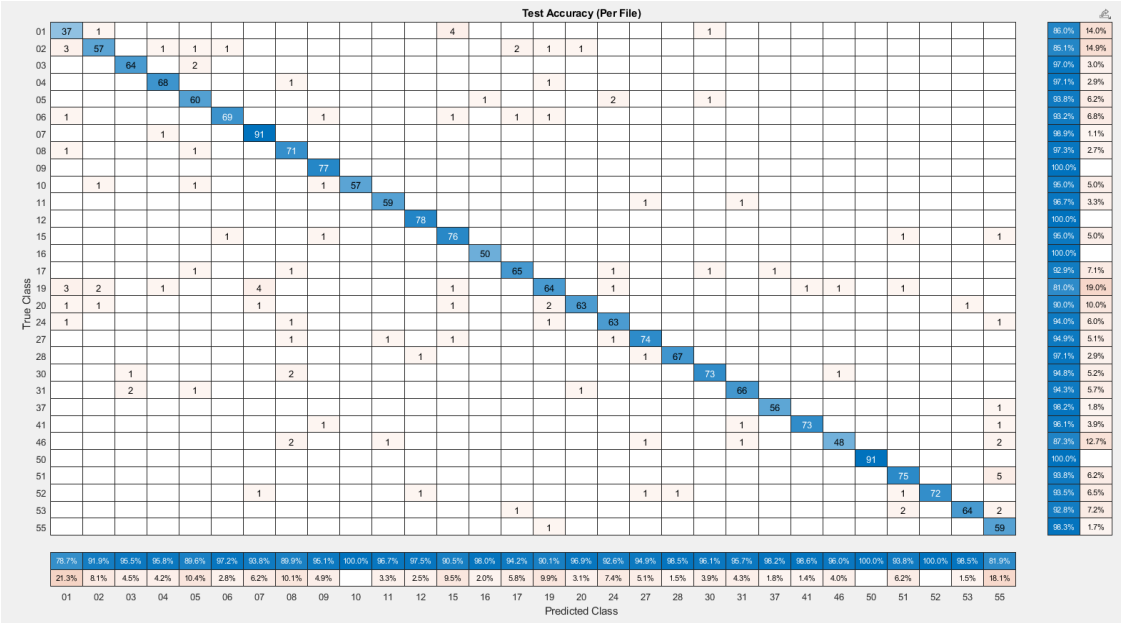

**Şekil 5.3.** SVM için Polinom kernel ile bulunan test konuşmacı doğruluk oranları

Aşağıdaki Tablo 5.5'te METUBET ve MNIST için KNN ile 14 ve 40 boyutlu öznitelik vektörleri ve K=3 ve K=5 komşuluk sayısına göre bulunan doğruluk oranları verilmiştir.

**Tablo 5.5.** METUBET ve MNIST için KNN ile bulunan doğruluk oranları

| MFCC+pitch sayısı | **METUBET** | | **MNIST** | |
|---|---|---|---|---|
| | **K=3** | **K=5** | **K=3** | **K=5** |
| **14** | 97.50 | 97.25 | 93.10 | 91.30 |
| **40** | 93.50 | 93.25 | 90.10 | 90.50 |

Ayrıca MNIST veri tabanı için SVM-polinom kernel ile konuşmacı başına test doğruluk oranları Şekil 5.4'te verilmektedir.




**Şekil 5.4.** MNIST için SVM polinom ile konuşmacı başına test doğruluk oranları

Şekil 5.4'teki doğruluk oranı 14 katsayı için %94.62 bulunmuştur. Tablo 5.6'de sonuçların tamamı MNIST için verilmiştir.

**Tablo 5.6.** MNIST için SVM ile bulunan konuşmacı doğruluk oranları

| MFCC+pitch sayısı | Polinom kernel | RBF kernel | Lineer kernel |
|---|---|---|---|
| 14 | 94.62 | 90.57 | 73.71 |
| 40 | 93.14 | 88.53 | 81.52 |

Tablo 5.7'de eğitim ve test için bir konuşmacıya ait tüm konuşma sinyallerinin çerçevelerine karşılık gelen yaklaşık bir ortak vektör kullanılarak konuşmacı belirleme yapılmıştır.

**Tablo 5.7.** METUBET için OVY ile bulunan konuşmacı doğruluk oranları

| MFCC+pitch sayısı | METUBET | MNIST |
|---|---|---|
| 14 | 95 | 96.67 |
| 40 | 100 | 96.67 |

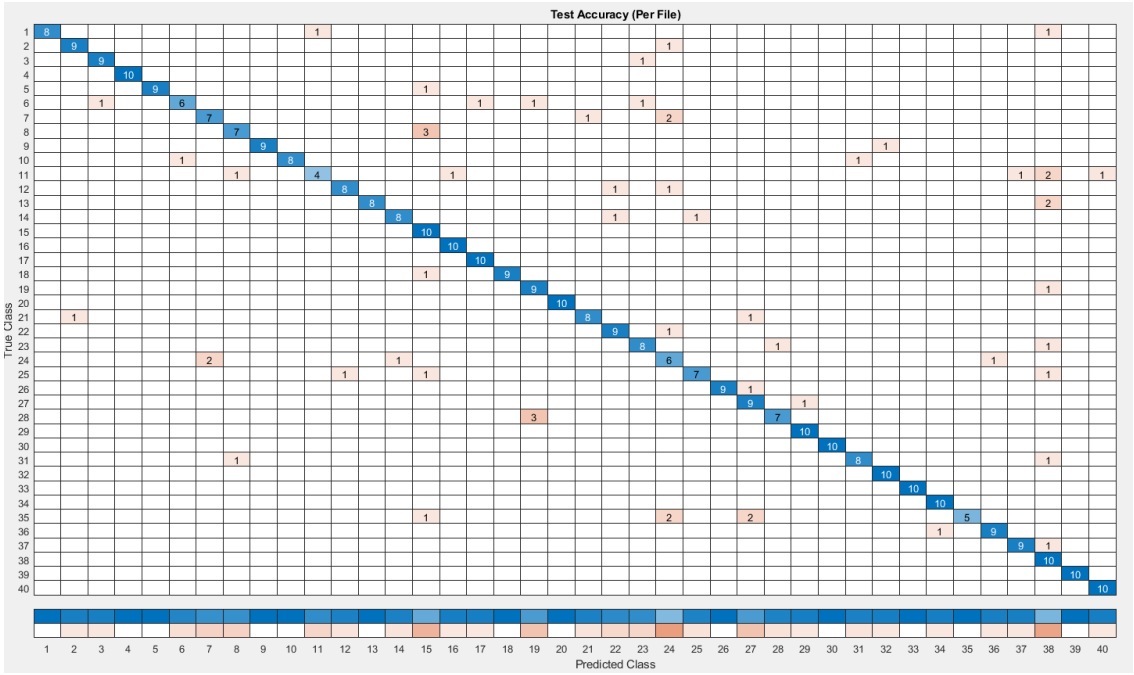
Yukarıdaki Tablo 5.7'den farklı olarak Tablo 5.8'de test aşamasında konuşmacının seslendirdiği bir ses sinyaline ait çerçeveler projeksiyon matrisiyle çarpılmış ve elde edilen vektörlerin ortalaması alınarak test ses sinyali için yaklaşık bir ortak vektör bulunmuştur.

Bulunan bu ortak vektör eğitim aşamasında elde edilen ortak vektörler ile karşılaştırılarak sınıflama yapılmıştır. OVY için Tablo 5.8'de METUBET ve MNIST için bulunan tanıma oranları verilmiştir.

**Tablo 5.8.** METUBET ve MNIST için OVY ile bulunan konuşmacı doğruluk oranları

| MFCC+pitch sayısı | METUBET | MNIST |
|---|---|---|
| 14 | 82.5 | 75.4 |
| 40 | 85.5 | 78.2 |

Aşağıdaki Şekil 5.5'te METUBET için bulunan tanıma oranları verilmiştir (%85.5).




**Şekil 5.5.** METUBET için OVY ile konuşmacı başına test doğruluk oranları

Çalışmada her konuşmacıya ait bir konuşma sinyalinin çerçevelerine karşılık gelen yaklaşık bir ortak vektör yerine yaklaşık ortak vektörlerden oluşan vektör dizileri kullanılarak hibrit OVY-KNN ve hibrit OVY-SVM sınıflayıcıları da kullanılmıştır. Elde edilen konuşmacı tanıma sonuçları Tablo 5.9, Tablo 5.10 ve Tablo 5.11'de verilmiştir.

**Tablo 5.9.** METUBET için hibrit OVY-SVM ile bulunan konuşmacı doğruluk oranları

| MFCC+pitch sayısı | Polinom kernel | RBF kernel | Lineer kernel |
|---|---|---|---|
| 14 | 84.5 | 83.75 | 74.25 |
| 40 | 90 | 89.25 | 79.50 |

**Tablo 5.10**. MNIST için hibrit OVY-SVM ile bulunan konuşmacı doğruluk oranları

| MFCC+pitch sayısı | Polinom kernel | RBF kernel | Lineer kernel |
|---|---|---|---|
| 14 | 82.56 | 81.21 | 74.46 |
| 40 | 86.12 | 83.25 | 76.67 |

**Tablo 5.11**. METUBET ve MNIST için hibrit OVY-KNN ile bulunan konuşmacı doğruluk oranları

| | METUBET | | MNIST | |
|---|---|---|---|---|
| MFCC+pitch sayısı | K=3 | K=5 | K=3 | K=5 |
| 14 | 87.50 | 88.00 | 83.24 | 84.57 |
| 40 | 88.50 | 89.50 | 86.62 | 86.62 |

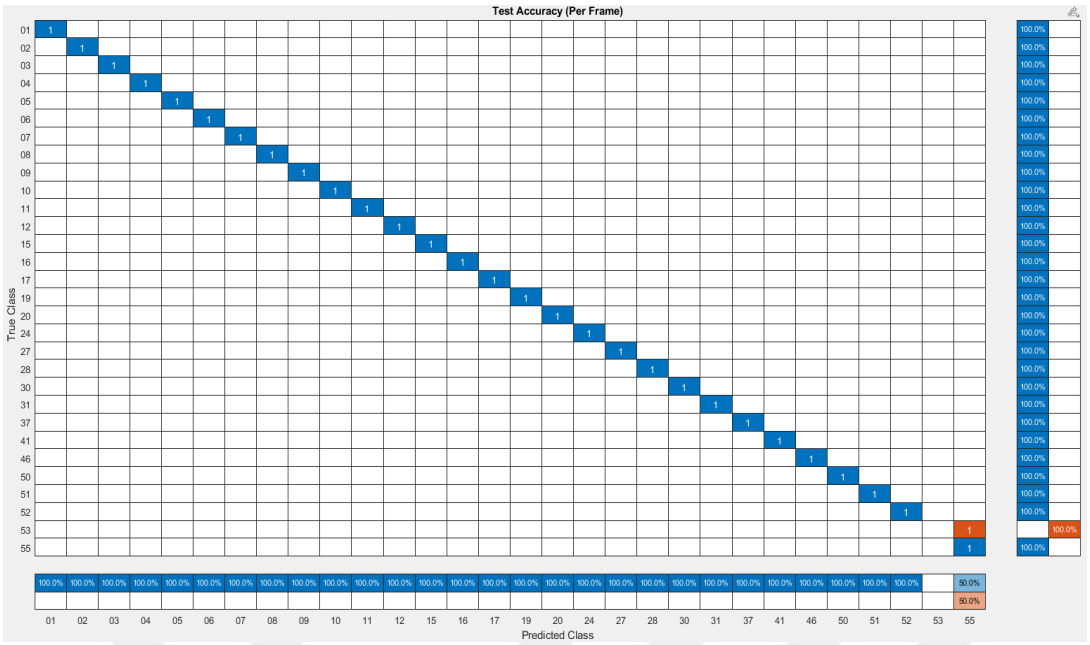
Yukarda verilen üç tablodaki değerlerin bulunma yaklaşımının aksine her konuşmacıya ait konuşma sinyallerinin çerçevelerine tümüne karşılık gelen bir adet ortak vektör kullanıldığında ise daha yüksek tanıma oranları elde edilmiştir. Ancak hibrit işlem için sadece SVM kullanılmış olup KNN için yapılmamıştır. Çünkü KNN'e uygulamak için bir konuşmacıya ait birden fazla ortak vektöre sahip olması gerekir. Bu sonuçlar Tablo 5.12 ve 5.13'te verilmiştir.

**Tablo 5.12**. METUBET için hibrit OVY-SVM ile bulunan konuşmacı doğruluk oranları

| MFCC+pitch sayısı | Polinom kernel | RBF kernel | Lineer kernel |
|---|---|---|---|
| 14 | 87.50 | 97.50 | 95.00 |
| 40 | 92.50 | 97.50 | 92.50 |

**Tablo 5.13.** MNIST için hibrit OVY-SVM ile bulunan konuşmacı doğruluk oranları

| MFCC+pitch sayısı | Polinom kernel | RBF kernel | Lineer kernel |
|---|---|---|---|
| 14 | 93.33 | 90.00 | 93.33 |
| 40 | 96.67 | 96.67 | 96.67 |




**Şekil 5.6.** Hibrit OVY-SVM için 40 katsayı ile MNIST için bulunan doğruluk oranları

Yukarıdaki Şekil 5.6'de bulunan sonuçlar test aşamasında bir konuşmacının tüm ses sinyallerinin çerçevelerinden sadece bir adet ortak vektör bulunarak yapılan sınıflamadır. Ancak bu durumun oluşması, eğer konuşmacının söylediği ses sinyalleri bir şekilde biriktirilip tanıma yapılması durumunda sağlanabilir. MNIST veri tabanı kullanılarak yapılan çalışmalardaki tek adetlik kelimelere ait ortak vektörler için ise %75 civarında bir tanıma başarımı elde edilmiştir. Yani tanıma başarımı düşmektedir.

# 6. SONUÇLAR VE ÖNERİLER

Çalışmalarda METUBET için en iyi ortalama konuşmacı tanıma oranları 40 katsayı için SVM polinom kernel ile %97.75 olarak ve OVY ile %100 bulunmuştur. OVY konuşmacı belirlemede kullanılmış ve çerçevelere ait öz niteliklerden elde edilen ortak vektörleri kullanarak yüksek bir tanıma oranı (%100) elde edildiği görülmüştür. KNN için K=3 ve 14 öznitelik ile bu oran %97.25 bulunmuştur. TDNN ile 20 MFCC katsayı kullanarak %92.75 bulunmuştur. MNIST için ise en iyi ortalama konuşmacı tanıma oranları TDNN ile 20 MFCC katsayısı kullanarak %96.17, 40 MFCC katsayısı kullanarak OVY ile %96.67, SVM polinom ile %94.63 ve KNN için K=3 ile %93.10 bulunmuştur. Sonuçlar incelendiğinde METUBET veri tabanı için özellikle OVY'nin 40 MFCC katsayı boyutunda en iyi sonucu verdiği (%100) görülmüştür. MNIST'te ise 14 boyutlu katsayı ile SVM polinom ile en iyi sonuçlar bulunmuştur. Sınıflandırıcılar içinde en iyi sonucu METUBET için OVY vermiştir.

Hibrit OVY-SVM ve Hibrit OVY-KNN sonuçları incelendiğinde gerek OVY, gerekse KNN ve SVM'ye göre daha düşük tanıma oranları elde edilmiştir. Bunun temel sebebi olarak her çerçeveye ait ortak vektörlerin sayısının fazla olmasıdır. Ancak bir konuşmacıya ait tüm konuşma sinyalini içeren çerçevelere ait bir adet ortak vektör bulunup hibrit yaklaşım uygulandığında METUBET ile 40 katsayı için OVY-SVM gaussian kernel ile %97.5 tanıma oranına erişilmiştir. Ayrıca MNIST ile 40 katsayı için OVY-SVM tüm kerneller ile %96.67 tanıma oranına erişilmiştir. Genel olarak en iyi sonuçların 40 katsayı ile elde edildiği görülmüştür.

## 7. KAYNAKLAR


Almaadeed, N., Aggoun, A., & Amira, A. (2015). Speaker identification using multimodal neural networks and wavelet analysis. IET Biometrics, 4(1), 18–28.

Al-Rawahy, S., Hossen, A., & Heute, U. (2012a). Text-independent speaker identification system based on the histogram of DCT-cepstrum coefficients. *International Journal of Knowledge-based and Intelligent Engineering Systems*, 16(3), 141–161.

An, N. N., Thanh, N. Q., & Liu, Y. (2019a). Deep CNNs with Self-Attention for Speaker Identification. IEEE Access apacoust.2019.107133

Calza, L., Gagliardi, G., Favretti, R. R., & Tamburini, F. (2020). Linguistic features and automatic classifiers for identifying mild cognitive impairment and dementia. Computer Speech & Language, 65, Article 101113.

Campbell, J. P. (1997). Speaker recognition: A tutorial. Proceedings of the IEEE, 85(9), 1437-1462.

Çeliktaş, H. (2019). Türkçe sesler ile konuşmacı kimliğinin doğrulanması/belirlenmesi (Master's thesis, Bursa Teknik Üniversitesi).

Çolak, R., & Akdeniz, R. (2018). Ses Sinyalinde Gürültü Saptama İçin Özgün Bir Yaklaşım. *European Journal of Engineering and Applied Sciences*, 1(1), 31-38.

Daqrouq, K. (2011). Wavelet entropy and neural network for text-independent speaker identification. Engineering Applications of Artificial Intelligence, 24(5), 796–802.

Dhakal, P., Damacharla, P., Javaid, A. Y., & Devabhaktuni, V. (2019). A Near Real-Time Automatic Speaker Recognition Architecture for Voice-Based User Interface. Machine Learning and Knowledge Extraction, 1, 504–520.

Disken, G., Tufekci, Z., Saribulut, L., & Cevik, U. (2017). A review on feature extraction for speaker recognition under degraded conditions. IETE Technical Review,34, 321–332.

Doddington, G. R. (1985). Speaker recognition—Identifying people by their voices. Proceedings of the IEEE, 73(11), 1651-1664.

Eskidere, Ö., & Ertaş, F. (2009), Bürünsel özelliklerin konuşmacı tanıma performansına etkisi. Uludağ Üniversitesi Mühendislik-Mimarlık Fakültesi Dergisi, 14-2.

Faragallah, O. S. (2018). Robust noise MKMFCC–SVM automatic speaker identification. *International Journal of Speech Technology*, 21(2), 185–192. https://doi.org/10.1007/s10772-018-9494-9

Fierrez, J., Morales, A., Vera-Rodriguez, R., & Camacho, D. (2018). Multiple classifiers inbiometrics. Part 1: Fundamentals and review. Information Fusion, 44, 57–64.

Furui, S. (1997). Recent advances in speaker recognition. Pattern recognition letters, 18(9), 859-872.

George, K. K., Kumar, C. S., & Panda, A. (2015). Cosine distance features for robust speaker verification. In Sixteenth annual conference of the international speech communication association.

Han, W., Chan, C. F., Choy, C. S., & Pun, K. P. (2006). An efficient MFCC extraction method in speech recognition. In 2006 IEEE International Symposium on Circuits and Systems (ISCAS) (pp. 4-pp). IEEE.

Jahangir, R., TEh, Y. W., Memon, N. A., Mujtaba, G., Zareei, M., Ishtiaq, U., Ali, I. (2020). Text-independent speaker identification through feature fusion and deep neural network. IEEE Access, 8, 32187–32202.

Jahangir, R., TEh, Y. W., Memon, N. A., Mujtaba, G., Zareei, M., Ishtiaq, U., … Ali, I. (2020). Text-independent speaker identification through feature fusion and deep neural network. IEEE Access, 8, 32187–32202.

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

Larcher, A., Lee, K. A., Ma, B., & Li, H. (2014). Text-dependent speaker verification: Classifiers, databases and RSR2015. Speech Communication, 60, 56–77.

Lawson, A., Vabishchevich, P., Huggins, M., Ardis, P., Battles, B., & Stauffer, A. (2011). Survey and evaluation of acoustic features for speaker recognition. In 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp.5444–5447): IEEE .

Liu, T., Das, R. K., Lee, K. A., & Li, H. (2022, May). MFA: TDNN with multi-scale frequency-channel attention for text-independent speaker verification with short utterances. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 7517-7521). IEEE.

Mallat, S. (1999). A wavelet tour of signal processing. Elsevier.

McLaren, M., Vogt, R., Baker, B., Sridharan, S., & Sridharan, S. (2010). Experiments in SVM-based Speaker Verification Using Short Utterances. In Odyssey (Vol. 17).

Rabiner, L., & Juang, B. H. (1993). Fundamentals of speech recognition. Prentice-Hall, Inc..

Saquib, Z., Salam, N., Nair, R. P., Pandey, N., & Joshi, A. (2010). A survey on automatic speaker recognition systems. In Signal Processing and Multimedia (pp. 134–145): Springer.

Sardar, V. M., & Shirbahadurkar, S. D. (2018). Speaker identification of whispering speech: An investigation on selected timbrel features and KNN distance measures. *International Journal of Speech Technology*, 21(3), 545–553.

Sarma, M., & Sarma, K. K. (2013). Vowel phoneme segmentation for speaker identification using an ANN-based framework. *Journal of Intelligent Systems*, 22, 111–130.

Shi, Y., Huang, Q., & Hain, T. (2020). Weakly Supervised Training of Hierarchical Attention Networks for Speaker Identification. arXiv preprint arXiv:2005.07817

Siam, A. I., El-khobby, H. A., Elnaby, M. M. A., Abdelkader, H. S., & El-Samie, F. E. A. (2019). A novel speech enhancement method using Fourier series decomposition and spectral subtraction for robust speaker identification. Wireless Personal Communications, 108, 1055-1068.

Soleymanpour, M., & Marvi, H. (2017). Text-independent speaker identification based on selection of the most similar feature vectors. International Journal of Speech Technology, 20(1), 99–108.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In Advances in neural information processing systems (pp. 3104–3112).

Tirumala, S. S., & Shahamiri, S. R. (2016). A review on Deep Learning approaches in Speaker Identification. In Proceedings of the 8th international conference on signal processing systems (pp. 142–147): ACM.

Tiwari, V. (2010). MFCC and its applications in speaker recognition. International journal on emerging technologies, 1(1), 19-22.

Tran, V.-T., & Tsai, W.-H. (2020). Speaker Identification in Multi-Talker Overlapping Speech Using Neural Networks. IEEE Access.

Umar, R., Riadi, I., Hanif, A., & Helmiyah, S. (2019). Identification of speaker recognition for audio forensic using k-nearest neighbor. Int. J. Sci. Technol. Res, 8(11), 3846-3850.

Vetterli, M., & Kovaˇcevic, J. (1995). Wavelets and subband coding: Prentice-Hall, Inc.

Wang, X., Xue, F., Wang, W., & Liu, A. (2020). A network model of speaker identification with new feature extraction methods and asymmetric BLSTM. Neurocomputing, 403, 167–181.

Wang, X., Xue, F., Wang, W., & Liu, A. (2020). A network model of speaker identification with new feature extraction methods and asymmetric BLSTM. Neurocomputing,403,167–181. https://doi.org/10.1016/j.neucom.2020.04.041

Wu, J.-D., & Lin, B.-F. (2009a). Speaker identification based on the frame linear predictive coding spectrum technique. Expert Systems with Applications, 36(4), 8056–8063.

Wu, J.-D., & Lin, B.-F. (2009b). Speaker identification using discrete wavelet packet transform technique with irregular decomposition. Expert Systems with Applications, 36(2), 3136–3143.

Wu, J.-D., & Tsai, Y.-J. (2011). Speaker identification system using empirical mode decomposition and an artificial neural network. Expert Systems with Applications, 38 (5), 6112–6117 .

Zhang, T., Shao, Y., Wu, Y., Geng, Y., & Fan, L. (2020). An overview of speech endpoint

## ÖZGEÇMİŞ

| KİŞİSEL BİLGİLER | |
|---|---|
| **Adı Soyadı:** | Esra GEZER |
| **Uyruğu:** | TC |
| **Orcid Numarası:** | 0000-0001-8570-5664 |

| EĞİTİM BİLGİLERİ | |
|---|---|
| **Lisans** | |
| **Üniversite:** | Erciyes Üniversitesi |
| **Fakülte:** | Mühendislığı Fakültesi |
| **Bölümü:** | Elektrik Elekronik Mühendisliği Bölümü |
| **Mezuniyet Yılı:** | 2007-2015 |
| **Yüksek Lisans** | |
| **Üniversite:** | Kırşehir Ahi Evran Üniversitesi |
| **Enstitü:** | Fen Bilimleri Enstitüsü |
| **Anabilim Dalı:** | İleri Teknolojiler Anabilim Dalı |
| **Mezuniyet Yılı:** | |
| **Doktora** | |
| **Üniversite:** | |
| **Enstitü:** | |
| **Anabilim Dalı:** | |
| **Mezuniyet Yılı:** | |
| | |

**Tezden Üretilen Makaleler ve Bildiriler**

**Uluslarası Hakemli Dergilerde Yayınlanan Makaleler**

**Uluslarası Konferans ve Sempozyumlarda Sunulan Bildiriler**

Gezer E., & Keser S., ‘’Speaker Identifcation Using SVM, KNN, and TDNN Classifers’’, 5. INTERNATIONAL MARMARA SCIENTIFIC RESEARCH AND INNOVATION CONGRESS, 2023 ANKARA/ TURKEY

**Ulusal Hakemli Dergilerde Yayınlanan makaleler**

……..

**Ulusal Konferans ve Sempozyumlarda Sunulan Bildiriler**

…….