DUDAK HAREKET ÖZELLİKLERİ KULLANILARAK TÜRKÇE KELİMELERİN SINIFLANDIRILMASI

Alper Yargıç

Yüksek Lisans Tezi

Bilgisayar Mühendisliği Anabilim Dalı

Ocak 2014

Bu tez çalışması Anadolu Üniversitesi Bilimsel Araştırma Projeleri Komisyonu Başkanlığı tarafından desteklenmiştir. Proje No: 1302F039

JÜRİ VE ENSTİTÜ ONAYI

Alper Yargıç'ın "Dudak Hareket Özellikleri Kullanılarak Türkçe Kelimelerin Sınıflandırılması" başlıklı **Bilgisayar Mühendisliği** Anabilim Dalındaki, Yüksek Lisans Tezi 20.01.2014 tarihinde, aşağıdaki jüri tarafından Anadolu Üniversitesi Lisansüstü Eğitim-Öğretim ve Sınav Yönetmeliğinin ilgili maddeleri uyarınca değerlendirilerek kabul edilmiştir.

		Adı Soyadı	**İmza**
Üye (Tez Danışmanı) :		**Yard. Doç Dr. MUZAFFER DOĞAN**	**……..**
Üye	**:**	**Prof. Dr. Ömer Nezih GEREK**	**……..**
Üye	**:**	**Yard. Doç. Dr. Sedat TELÇEKEN**	**……..**

Anadolu Üniversitesi Fen Bilimleri Enstitüsü Yönetim Kurulu'nun …….. tarih ve …….. sayılı kararıyla onaylanmıştır.

Enstitü Müdürü

ÖZET

Yüksek Lisans Tezi

DUDAK HAREKET ÖZELLİKLERİ KULLANILARAK TÜRKÇE KELİMELERİN SINIFLANDIRILMASI

Alper YARGIÇ

Anadolu Üniversitesi
Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

Danışman: Yard. Doç. Dr. Muzaffer DOĞAN

2014, 52 Sayfa

İşitme engellilerin ses terapisinde sesin yanında dudak okuma verileri de kullanılmaktadır. Dudak okuma üzerine literatürde çok sayıda çalışma olmasına rağmen, tümleşik kızılötesi sensörü sayesinde derinlik bilgisini de ölçebilen MS Kinect kamerasını kullanarak Türkçe kelimeler için yapılmış başka bir çalışmaya rastlanamamıştır. Bu çalışmanın amacı, MS Kinect kamerası kullanılarak sık kullanılan bazı Türkçe kelimelere ait, derinlik bilgisini de içeren görsel bir veri seti hazırlamak ve bu veri seti üzerinde en iyi sınıflandırma yöntemini araştırmaktır. Sınıflandırma için Yapay Sinir Ağları, KNN ve Dinamik Zaman Bükmesi gibi yöntemler kullanılmıştır. Ayrıca, işitme engelli çocukların ve konuşma terapisine ihtiyaç duyan kişilerin, eğitmen yardımıyla dudak hareketlerini taklit yeteneklerini geliştirmeyi hedefleyen ve iki telaffuzu karşılaştıran bir yazılım geliştirilmiştir. Bu yazılım ve oluşturulan veri seti, işitme engelli çocuklar için eğitim materyali olarak kullanılabilecektir.

Anahtar Kelimeler: Dudak Okuma, MS Kinect Kamera, 3B Yüz Tanımlama, Yapay Sinir Ağları, Dinamik Zaman Bükmesi, Konuşma Notlandırma

ABSTRACT

Master of Science Thesis

CLASSIFICATION OF TURKISH WORDS BY USING LIP MOTION FEATURES

Alper YARGIÇ

Anadolu University
Graduate School of Sciences
Computer Engineering Program

Supervisor: Assist. Prof. Dr. Muzaffer DOĞAN

2014, 52 pages

Information obtained from Lip Reading in addition to voice data is used in voice therapy of hearing-impaired persons. Although there are many studies on the applications of Lip Reading, no study exists on the recognition of Turkish Words using MS Kinect camera, which has a built-in integrated infrared sensor that measures the depth information. The aim of this project is to construct a data set on frequently used Turkish words containing depth information, investigating the best lip reading classification method on this data set. For classification, techniques such as Artificial Neural Networks, KNN, and Dynamic Time Warping has been used. Furthermore, a complementary software was developed, which improves the lip imitation skills of hearing-impaired children and people who need speech therapy with the aid of an instructor by comparing two pronunciations. The developed software and the generated data set will be used as educational materials for hearing-impaired children.

Keywords: Lip Reading, MS Kinect Camera, 3D Face Tracking, Artificial Neural Networks, Dynamic Time Warping, Speech Scoring.

TEŞEKKÜR

Bu çalışmada bana yol gösteren danışmanım Sayın Yard. Doç. Dr. Muzaffer Doğan'a, bu konudaki bilgi ve birikimlerini benimle paylaşan Sayın Yard. Doç. Dr. Alper K. UYSAL'a ve Sayın Araş. Gör. Sevcan YILMAZ'a, veri seti oluşturmamda bana yardımcı olan Anadolu Üniversitesi Bilgisayar Mühendisliği Bölümü öğrencilerine ve çalışma arkadaşlarıma ayrıca tez çalışmam sırasında bana maddi manevi her türlü desteğini sunan aileme teşekkür ederim.

Alper YARGIÇ

Ocak – 2014

İÇİNDEKİLER

ŞEKİLLER DİZİNİ

ÇİZELGELER DİZİNİ

SİMGELER ve KISALTMALAR DİZİNİ

ANN	:	Yapay Sinir Ağları (Artificial Neural Network)
CALL	:	Bilgisayar Destekli Dil Öğrenme Sistemini (Computer Assisted Language Learning System)
CAPT	:	Bilgisayar Destekli Telaffuz Eğitimi (Computer Assisted Pronunciation Training)
DCT	:	Ayrık Kosinüs Dönüşümü (Discrete Cosine Transform)
DTW	:	Dinamik Zaman Bükmesi (Dynamic Time Warping)
DWT	:	Ayrık Dalgacık Dönüşümü (Discrete Wavelet Transform)
FPS	:	Saniyedeki Çerçeve Sayısı (Frame Per Second)
FTE	:	Yüz İzleme Motoru (Face Tracking Engine)
HMM	:	Saklı Markov Model (Hidden Markov Model)
IR	:	Kızıl Ötesi (Infrared)
IWR	:	İzole Edilmiş Kelimeleri Sınıflandırma (Isolated Word Recognition)
KNN	:	K-En Yakın Komşu (K-Nearest Neighbour)
NUI	:	Kullanıcı Arayüzü (Natural User Interface)
PCA	:	Temel Bileşen Analizi (Principle Component Analysis)
RPROP	:	Esnek Geri Yayılım Algoritması (Resilient Backpropagation)
SDK	:	Yazılım Geliştirme Kiti (Software Development Kit)
SVM	:	Destek Vektör Makinesi (Support Vector Machine)

1. GİRİŞ

Ses bilgisi, dudak okuma sistemlerinde sınıflandırma başarısını arttırmak için konuşma esnasındaki ağız hareketlerini tanımlayıcı ve destekleyici bir kaynaktır. Konuşma sürecinde meydana gelen dudak hareketleri gibi görsel özelliklerin önemi göz ardı edilemez [1,2]**.** Günlük hayatta, konuşma esnasında, normal duyma yetisine sahip insanlar ve işitme kaybı olan insanların duydukları sesleri anlamlandırmak, aynı zamanda anlamayı kolaylaştırmak için görüntüden elde ettikleri verileri kullandıkları bilinen bir gerçektir [2].

İşitme engelli insanlar konuşma esnasında görsel verileri kullanarak konuşan kişiyi anlamaya çalışmaktadırlar. Sadece görüntüden elde edilen veriler kullanılarak konuşmacının ağzından çıkan bütün kelimelerin algılanması mümkün değildir. Çünkü konuşma esnasında oluşan bütün sesler sadece dudağın hareketleri ile meydana gelmez. Konuşma esnasında sesin oluşumu ses telleri ile başlar, ses yolunda, yutakta, ağızda ve burunda şekillenir. Diş, dil, sert ve yumuşak damak ile dudak kullanılarak ses şekil değiştirir [3]. Bunun yanı sıra bazı fonemlerin telaffuz biçimi görsel olarak birbirine benzerlik gösterse de, oluşan ses farklı fonemlere karşılık gelebilir [4]. İşitme engelli kişi, konuşma esnasında yalnızca dudak hareketlerini izlemez; bununla birlikte yüzdeki ifadeleri el ve vücuttaki hareketleri de konuşmayı algılamada ipucu olarak kullanır. Bu nedenle dudak okuma işlemi sırasında ses verisi kullanılmadan bütün kelimelerin birebir anlaşılması mümkün değildir.

Gelişen teknoloji ile birlikte konuşma tanıma sistemleri ile ilgili çalışmalar yoğunluk kazanmıştır. Geliştirilen bu sistemler, konuşma bozuklukları olan insanların telaffuz kalitesini arttırmak, yeni bir dil öğrenmeye çalışanların telaffuz kalitesini belirlemek ve insanların dudak okuma becerilerini geliştirmek için kullanılmaktadır [5,6,7,8].

Konuşmanın algılanmasında ses ile birlikte görüntüden elde edilen bilgilerin ayırt edici özellikleri vardır. Yalnızca ses ya da yalnızca görüntü ile elde edilen verilerle yapılan sınıflandırmalara göre, ses ve görüntü verisinin birlikte kullanıldığı veri setleri ile yapılan sınıflandırmalar daha başarılıdır [9]. Dudak okuma, dinleyicinin görme duyusu tarafından algılanan bilgileri yorumlayıp

konuşma olarak algılama becerisidir. İnsanların eğitim ile gerçekleştirebildiği bu tekniği bilgisayar yardımı ile elde etmek mümkündür.

Görsel veriler ile yapılan sınıflandırmalar *şekil tabanlı* (shape-based) ve *görünüm tabanlı* (appearance-based) olarak iki temel kategoriye ayrılabilir [10]. Şekil tabanlı sistemlerde ağız ve dudak üzerinde önceden tanımlanmış noktaların konumları öznitelik değerleri olarak kullanılır. Örneğin, dudağın yükseklik ve genişlik parametreleri bu tür özniteliklerdendir [9]. Görünüm tabanlı sistemlerde ağız alanı çevresindeki imgenin piksel yoğunluk değerleri temel alınır [11].

Dudak okuma esnasında kafa hareketleri sistemin başarısını etkilemektedir. İki boyutlu imgeler üzerindeki noktalar arasındaki uzaklıklar, kafanın sağa-sola oynatılmasıyla değişebilmektedir. Bu yüzden derinlik bilgisini kullanarak dudak okuma sistemleri de geliştirilmiştir [12]. Noktalar arasındaki uzaklıklar yerine noktalar arasında kalan açıların öznitelik olarak kullanıldığı çalışmalar da mevcuttur. Iwano ve ark., video görüntülerinin yanında ses özniteliklerini de kullanarak dudak okuma başarısını arttırmışlardır [13,14].

Farklı kişilerden alınan görüntülerde aynı kelimenin telaffuzu farklı sürelerde gerçekleşebilmektedir. Başarılı bir tanıma için sinyallerin özelliklerini kaybetmeden aynı uzunluğa getirilmesi önem kazanmaktadır. Bu problemin üstesinden gelmek için Myers ve ark. [15], Dinamik Zaman Bükme *(Dynamic Time Warping-DTW)* algoritması kullanmışlardır. Genelleştirme ve sınıflandırma performanslarının arttırılması amacıyla Yapay Sinir Ağlarını *(Artificial Neural Network-ANN)* [16,17] ve Destek Vektör Makinelerini *(Support Vector Machines-SVM)* [18] kullanan çalışmalar da mevcuttur. Eğitim için fazla veriye ihtiyaç duyan ve veri setine yeni kelime eklendiğinde yeniden eğitilmesi gereken bu son iki yöntem yerine dudak okuma sistemleri için K-En Yakın Komşu *(K-Nearest Neighbor-KNN)* sınıflandırıcısı yöntemi de tercih edilmektedir [19]. Shin ve ark. [19], navigasyon cihazı üzerinde kullanılmak üzere Kore dilindeki navigasyon ile ilgili kelimelerin hem ses hem de görüntü bilgisini analiz eden, sınıflandırma için Saklı Markov Modeli *(Hidden Markov Model-HMM)*, KNN ve ANN yöntemlerini kullanan bir sistem geliştirmiştir.

Puviarasan ve ark. [20], Ayrık Dalgacık Dönüşümü *(Discrete Wavelet Transform-DWT)* ve Ayrık Kosinüs Dönüşümü *(Discrete Cosine Transform-DCT)*

kullanarak görüntü verisinden öznitelikleri ayrıştıran ve bu öznitelikleri HMM ile sınıflandıran, Hintli işitme engelli kişiler için bir dudak okuma yöntemi geliştirmişlerdir.

Konuşma terapisinde kullanılmak üzere yalıtılmış Türkçe fonemleri ve sözcükleri içeren diğer bir veri seti ise Türk ve ark. [21] tarafından oluşturulmuştur. Bu çalışmada, söylenişleri birbirine yakın olan sesler için sınıflandırma başarımlarını test etmek için HMM yöntemi kullanılmıştır. Çetingül ve ark. [22], dudak hareket özniteliklerini kullanarak Türkçe kelimeler üzerinde bir sınıflandırma çalışması yapmışlar ve görüntünün yanında ses verisinin de kullanılmasıyla sadece görüntünün kullanıldığı duruma göre başarımın arttığını göstermişlerdir. Dudak hareketleriyle birlikte sesi de kullanan başka bir çalışma da, Kaynak ve ark. [23] tarafından yapılmıştır. Bu çalışmada, dudağın geometrik görsel öznitelikleri araştırılmış ve sınıflandırma işleminde iki dudak arasındaki açıklığın ve dudak kenar açılarının en önemli öznitelikler olduğu sonucuna varılmıştır. Galatas ve ark. [24], MS Kinect kamerasından alınan ve derinlik bilgisine sahip görüntüler üzerinde hem görüntü hem de ses verisini kullanarak bir dudak okuma çalışması yapmışlardır.

Görüntüden elde edilen verilerin önemi gürültülü ortamlarda daha da artmaktadır. Düşük gürültü seviyesine sahip ortamlarda yapılan uygulamalarda sadece ses verisi kullanılarak yapılan sınıflandırmalar %95'in üzerinde sınıflandırma başarısına sahiptir [19]. Literatürde yapılan araştırmalarda sadece görsel veriler kullanarak sınıflandırma yapılan sistemlere de rastlanmıştır [25,26].

Dudak okuma uygulamaları, konuşma telaffuz kalitesi değerlendirme sistemleri ve bilgisayar destekli dil eğitim sistemlerinde yaygın olarak kullanılmaktadır. Bilgisayar destekli dil eğitimi önceden tanımlanmış bir metin ya da görüntüye dayalı egzersizler ile kullanıcıya dil eğitiminde yardımcı olmayı amaçlar. Ayrıca konuşma sesinin veri olarak kabul edilmesiyle pratik hale getirilmişlerdir. Bu sistemlerde, yazılım bir eğitmen gibi kullanıcının kelime telaffuz kalitesini değerlendirerek, kullanıcıya geri bildirim sağlar [6].

Zhang ve ark. [27], İngilizce kelimelerin telaffuz kalitesini, bilgisayar yazılımı ile otomatik değerlendirecek bir sistem geliştirmiştir. Sözlü okuma ile elde edilen sesli verileri, okuma bütünlüğünü içerecek şekilde analiz etmiştir.

Manuel puanlama kurallarına göre, telaffuz ve akıcılık özniteliklerini içeren bir dizi çıkarılmıştır. Elde edilen öznitelikler ile SVM regresyonu kullanılarak sınıflandırma yapılmış ve telaffuz kalitesi belirlenmiştir.

Cincarek ve ark. [7], konuşmacılar için telaffuz kalitesini otomatik olarak değerlendirecek bir yaklaşım geliştirmişlerdir. Cümleler ve kelimeler skor birimleri olarak kabul edilmiştir. Ayrıca, yanlış telaffuz ve fonem karışıklıkları için hedef dil fonem seti açıklamaları türetilmiştir. Yanlış telaffuz edilmiş bir kelimenin algılanmasında ve kelimeleri sınıflandırmada HMM kullanılmış olup, değerlendirme işlemi sözcük düzeyinde yapılmıştır. Önerilen yöntem Bilgisayar Destekli Telaffuz Eğitimi *(Computer Assisted Pronunciation Training-CAPT)* için bir sistemin puanlama modülünün bir parçası olarak kullanılabilir. Örüntü ve konuşma tanıma yöntemleri cümle ve kelime düzeyinde puanlama için uygun öznitelik setleri geliştirmek için uygulanır. Doğrusal öznitelik kombinasyonu ile elde edilen skor, insan referansı ile elde edilenden daha düşük korelasyona sahiptir. Kelime bazında değerlendirmeye en uygun model Markov Zincir Modelidir. Bu modelde ses birimlerinden bir veya daha fazlasında yapılan yanlışlar, belirgin bir telaffuz hatası olarak kabul edilir.

Bilgisayar destekli dil eğitim sisteminin öğrenci ve öğretmen için birçok potansiyel faydaları vardır. Bu sistemler öğretmene ihtiyaç duymadan öğrenciye sürekli geri bildirim sağlamakta ve böylece bireysel çalışmalar kolaylaşarak ezberci öğrenme yerine interaktif kullanım teşvik edilmiş olmaktadır. Bilgisayar destekli sistemler değerlendirme işlemlerini kolaylaştırmak için de kullanılır.

Witt ve ark. [5], HMM ile konuşma tanıma sistemi çerçevesinde otomatik telaffuz değerlendirmesi için akustik olasılık tabanlı yöntemleri incelemektedir. Bilgisayar Destekli Dil Öğrenme Sisteminin *(Computer Assisted Language Learning System-CALL)* etkili olabilmesi için, hataların hemen düzeltilmesi gerekmektedir. Ayrıca daha uzun bir dil yetkinlik geri bildirimini sağlamak için, telaffuz kalitesini tutarlı bir şekilde ölçülmesi önemlidir.

Mevcut otomatik telaffuz puanlama sistemleri esas olarak kelime ve sözcük parçası tonlama, vurgulama, sesteki vurgu ve ritme odaklanır. Bu sistemler tipik olarak; eğitim materyali içerisinde kullanılan her kelime için bir modele ihtiyaç duyarlar. Bu nedenle metne bağımlı olarak çalışırlar ve eğitim materyali

üzerindeki herhangi bir değişiklikte sistemi eğitmek için yeni verilere ihtiyaç duyarlar. Sistem seçilen ses-bilimsel *(phonemic)* hataları bulmayı ve öğretmeyi amaçlar. HMM kullanılarak oluşturulan konuşma tanıma sistemleri, telaffuz başarımını ölçmek ve değerlendirmek için bütün cümle yerine, cümlenin daha küçük parçaları kullanılır [5,28,29].

Arias ve ark. [30], ses tonlama ile elde edilen verilerin niteliği ve bu verilerin ikinci dil öğrenme işlemi için uygunluk düzeyini incelemişlerdir. Elde edilen veriler, fonetik kurallara göre kolayca "doğru" veya "yanlış" olarak sınıflandırılabilir. Söyleyişteki tonlama-ses uyumu; duygular, niyet ve düşünceler hakkında bilgi içerir. Sonuç olarak, tonlamayı-ses uyumunu doğru ya da yanlış olarak sınıflandırmak yerine, öğrenciyi belirlenen bir referans tonlama modelini takip etmesi için motive etmek daha anlamlı bir çalışmadır. Öğrenci tarafından söylenen bir ifade-ses-cümle vs. doğrudan bir referans ile karşılaştırılmıştır. Tonlama ve enerji eşyükselti eğrileri *(contour)* ile benzerlikleri DTW ile karşılaştırılır. Önerilen yöntem; tonlama değerlendirme prosedürünü, öğrencinin kelimeyi telaffuz kalitesini belirlemeye çalışır. Verilen referans ifadeyi dinledikten sonra, öğrenci verilen referans tonlama eğrisini takip etmeye çalışır. Daha sonrasında referans ve test için kullanılan ifadeler DTW kullanılarak hizalanır. Test ve referans ifadeleri için ayrı ayrı ses perdesi yakalama *(Pitch Detection)* uygulanmıştır. Ses perdesi yakalama, periyodik sinyaller ya da konuşma gibi bir dijital kaydın perde ya da temel frekans aralığını tahmin etmek için kullanılan bir algoritmadır. Referans ve test tonlama şablonları arasındaki eğilim benzerliği *(trend similarity)* DTW hizalama algoritması kullanılarak çerçeve-çerçeve değerlendirilir. Temel frekans *(fundamental frequency)*, bir periyodik dalganın en düşük frekansı olarak tanımlanır. Referans ve test verisinin temel frekansı arasındaki eşyükselti eğrisinin farklılığını çerçeve-çerçeve karşılaştırmak yerine, iki eğri arasındaki korelasyonu hesaplanmıştır. Son olarak, hece vurgusu *(syllable stress)*, tonlama eğrisi *(intonation curve)* ve çerçeve enerjisinin kombinasyonundan elde edilen veriler kullanılarak telaffuz kalitesi değerlendirilmiştir. Tonlama eğrisi, bir kelimeyi söylerken, söyleyişte yükselip düşen ses perdesidir. Önerilen sistem metne bağımlı değildir, yani referans olarak

kullanılacak bir söyleyişe ihtiyacı yoktur. Kullanılan bu yöntemler ile öğrencinin telaffuzu içerisinde ortaya çıkan fonetik kalitenin etkisi en aza indirilmiş olur.

Türk ve ark. [21], yaptıkları çalışmada konuşma terapisine yönelik, konuşma tanıma yöntemlerini araştırmışlardır. Yapılan çalışmada, Türkçe'de kullanılan izole edilmiş kelimeleri içeren bir veri seti oluşturmuşlardır. Oluşturulan veri seti, sesten elde edilen akustik verilerden meydana gelmektedir ve HMM ile modellenmektedir. Oluşturulan veri setinde kullanılan akustik öznitelikler; mel frekansı spektrum katsayıları, enerji ve sessizlik olasılığı ile bunlara karşılık gelen farklılıklar ve ivme parametreleridir. HMM modeli eğitimi *Boum-Welch* algoritmasıyla gerçekleştirilmiş, tanıma işlemi için ileri-yön *(Forward)* algoritması kullanılmıştır.

Kumar ve ark. [31], yaptıkları çalışmada kişinin profil ve ön cephe görünüşünün, dudak okumadaki etkisini araştırmışlar ve sınıflandırmada kullanılmak üzere öznitelik çıkarımını gerçekleştirmişlerdir. Profil görüntüsü için dudak çıkıntısı ve dudak kontörleri arasındaki yüksekliği öznitelik olarak kullanmışlardır. Ön cephe görüntüsü için 3 öznitelik kullanmışlardır. Bunlar dudağın orta noktası ile üst ucu arasındaki yükseklik, alt ucu ile orta nokta arasındaki yükseklik ve dudağın açıklık genişliğidir. Sonuç olarak yan profilden elde edilen öznitelikler kullanılarak yapılan sınıflandırmada daha düşük hata oranını yakalamışlardır.

1.1.Problemin Tanımı ve Amaç

Literatürde yapılan araştırmalar sonucunda görüntüden elde edilen verilerin; konuşmada kullanılan ses birimlerini, fonemleri ve kelimeleri sınıflandırmak için kullanılan önemli bir öznitelik olduğu anlaşılmıştır. Ayrıca konuşma terapisi ve telaffuz değerlendirme uygulamaları için de görüntüden elde edilen verilerin kullanımı önemli ve ayırt edici bir özniteliktir. Görüntüden elde edilen veriler ile konuşma terapisinde ve kişilerin dudak okuma becerilerini geliştirmede kullanılabilecek bir sistemin tasarlanması planlanmıştır.

Mevcut sistemler genel olarak ses verisi kullanarak bir sınıflandırma ve değerlendirme işlemi yapmaktadır. Bu sistemler kelimenin bütünü veya fonemlere

göre doğru ya da yanlış şeklinde sonuçlar üreterek değerlendirme işlemini gerçekleştirirler.

İşitme engelliler, dudak okuma ile konuşulanı anlayabilmekte ve hatta karşıdaki kişiye sesli olarak cevap verebilmektedir. İşitme engellilerin sesleri daha düzgün çıkartabilmeleri için dudak hareketlerini doğru bir şekilde taklit etmesi gerekir. Böylece dudak hareketleri taklidindeki başarı oranı arttırılarak işitme engellilerin hem birbirleriyle hem de normal duyma yetisindeki insanlarla iletişim başarısı arttırılabilecektir. Geliştirmeyi tasarladığımız sistemde işitme engelli insanların dudak hareketlerini tutarlı biçimde taklit etmelerini hedeflemekteyiz. Bu bağlamda işitme engelli insanların konuşma esnasında ses çıkarma yetenekleri olmadığı için görsel-işitsel sistemlerin tutarlılığı ve sınıflandırma başarısı yüksek olsa da sadece görüntü verisi kullanılarak sınıflandırma yapılacaktır. Bu sistem ayrıca normal duyma yetisine sahip olup, konuşma bozuklukları olan kişilerin konuşma telaffuz kalitesini de arttırmayı hedeflemektedir.

Bu çalışmada, işitme engellilerin ve konuşma terapisine ihtiyaç duyan kişilerin, dudak hareketlerini kendi kendilerine analiz etmelerini sağlamak amacıyla bir dudak okuma uygulaması geliştirmesi planlanmaktadır.

Mevcut sistemler genel olarak kişiye bağımlı olarak çalışmaktadır ve bununla birlikte önceden belirlenmiş kelime ya da cümlelerden oluşan bir veri seti ile eğitilmeleri gerekmektedir. Önceden belirlenmiş kelime ya da cümle kalıpları dışında değerlendirme işlemi gerçekleştirmediği için kelimelerin yanı sıra tasarlandığı dil kalıplarına da bağımlı kalması gerekmektedir. Yapılan çalışmada, kişiden bağımsız olarak ve herhangi bir eğitim verisine ihtiyaç duymadan kişilerin konuşma esnasındaki dudak hareketlerinden söylenen kelimelerin birbirlerine benzerliklerini çıkarmak için kullanılacak öznitelikler oluşturulmaya çalışılmıştır. Oluşturulan sistemde öğretici herhangi bir metne bağımlı kalmadan anlık olarak belirlediği bir kelimeyi kamera karşında söyleyecek ve söylediği bu kelimeyi kullanıcı kamera karşısında tekrar ederek iki konuşma arasındaki benzerlik çıkarımı yapılacaktır.

Bu işlemin gerçekleşmesi için, sistemin tutarlığını test edecek bir veri seti oluşturulup; bu veri seti kullanılarak ayırt ediciliği en yüksek olan özniteliklerin belirlenmesi gerekmektedir. Veri setini oluşturmak için tasarlanan yazılım MS

Kinect kamerası tarafından elde edilen görüntüleri analiz ederek, görsel veriyi sayısal veriye çevirip kayıt altına almaktadır. Kaydedilen veri setinin ve kayıt sisteminin tutarlılığı KNN ve ANN ile test edilerek sistemin sınıflandırma başarısı değerlendirilmiştir. Sınıflandırma işlemindeki ayırt edici özniteliklerin çıkarımı Temel Bileşen Analizi *(Principle Component Analysis-PCA)* ile gerçekleştirilmiştir. Bu öznitelikler kullanılarak kelimenin telaffuz kalitesi DTW yöntemi ile belirlenmiştir.

1.2.Tezin Ana Hatları

Bölüm 2'de tez çalışmada görüntü elde etmek için kullanılan Microsoft Kinect kamerası ve bu kamera ile entegre çalışan Microsoft Kinect yazılım geliştirme kiti hakkında bilgi verilmiştir. Bu bölümde sınıflandırma işlemleri için kullanılan KNN yöntemi ve yapay sinir ağlarının genel yapısı ile geri yayılım algoritması ile ilgili açıklamalar da yer almaktadır. Bölümün sonunda ise iki kelimenin birbirine olan benzerliğini test etmek için kullanılan DTW yöntemi hakkında bilgi verilmiştir. Bölüm 3'te kaydedilen görüntüden veri setini elde etmek için yapılan anlık enerji hesabı, kelime başlangıç bitiş noktasının belirlenmesi, kübik şerit interpolasyonu ile yapılan normalizasyon işlemi ve temel bileşen analizi ile veri setinin boyutunu indirgeme hakkında bilgi verilmiştir. Bölüm 4'te KNN ve yapay sinir ağı ile elde edilen veri setinin sınıflandırma başarısı test edilmiştir. Elde edilen sonuçlar doğrultusunda kelime telaffuzunu ifade eden en iyi açılar bulunmuştur. Bu açı değerleri ve açı değerlerinden elde edilen öznitelikler kullanılarak dinamik zaman bükmesi yöntemi ile iki konuşmanın birbirine olan benzerliğini belirlemek için bir yöntem önerilmiştir. Bölüm 5'te ise elde edilen sonuçlar yorumlanmıştır.

2. KINECT KAMERASI ve KULLANILAN YÖNTEMLER

Konuşma tanıma ve telaffuz kalitesi değerlendirme sistemlerinde, konuşmacıdan tutarlı veriler toplamak ve bu verileri anlamlı şekilde kayıt altına almak oldukça önemlidir.

Bu çalışmada, konuşmacıdan elde edilen verilerin anlamlı olarak toplanması ve kayıt altına alınması için Microsoft tarafından geliştirilen MS Kinect kamerası ve kamera ile entegre çalışan MS Kinect Yazılım Geliştirme Kit'i *(MS Kinect Software Development Kit-SDK)* kullanılmıştır. Yapılan çalışmada Kinect kamerasının tercih edilme nedeni, 2 boyutlu görsel veriler ile birlikte belirlenen noktaların derinlik bilgisinin de kullanılacak olmasıdır.

2.1.MS Kinect Kamerası

MS Kinect kamerasının sahip olduğu bileşenler Şekil 2.1'de gösterilmiştir. Bu bileşenler aşağıdaki gibi sıralanabilmektedir:

- RGB (Red-Green-Blue) kamerası
- Kızılötesi-IR emitörü
- Kızılötesi-IR derinlik sensörü
- Eğim motoru
- Mikrofon dizisi
- Led

Şekil 2.1. MS Kinect kamerası ve bileşenleri

RGB Kamerası: Bu kamera renkli video verilerini yakalamak ve aktarma işleminden sorumludur. Kameranın işlevi kaynaktan gelen kırmızı, yeşil ve mavi renkleri yakalamaktır. Kinect kamera 640x480 piksel çözünürlükle 30 fps. ya da 1280x960 piksel çözünürlükle 12 fps. görüntü akışını destekler. Kamera; 43 derece dikey eksende, 57 derece yatay eksende aktif görüntü çekim aralığına sahiptir.

IR Yayıcı (Emitör) ve Derinlik Sensörü: Kinect derinlik sensörü IR emitörü ve IR derinlik sensöründen oluşmaktadır. Kamera karşısındaki nesnelerin derinlik bilgilerine ulaşmak için bu iki donanımın birlikte çalışması gerekmektedir. IR yayıcısı karşısındaki nesnelere kızıl ötesi ışınlar yansıtan bir projektöre sahiptir. Bu projektör tarafından yayılan kızıl ötesi ışınlar, derinlik sensörü aracılığı ile toplanır. Böylece sensör ile karşısındaki nesne arasındaki mesafe bilgisi elde edilir.

2.2.Kinect Yazılım Geliştirme Kiti

Kinect Yüz İzleme Yazılım Geliştirme Kiti *(Kinect Face Tracking SDK)* gerçek zamanlı olarak yüzü tanımlayıp takip edebilen uygulamalar geliştirmeye olanak sağlar. Yüz İzleme Motoru *(Face Tracking Engine-FTE)*; Kinect kamerası tarafından elde edilen görüntüleri renk ve derinlik bilgisini kullanarak analiz eder ve kafanın pozisyonunu, yüzdeki ifadeleri tahmin eder [32].

Yüz izleme motoru yüzün şeklini oluşturmak için renk ve derinlik bilgisine ihtiyaç duyar. Yüzün şekli oluşturulduktan sonra önceden belirlenmiş noktalar için pozisyon bilgilerine ulaşılabilir. Bu işlem oldukça tutarlı sonuçlar vermekle birlikte işlemci için yoğun bir süreçtir [32].

Derinlik değerleri için *yakın* veya *standart* çekim modu olmak üzere iki tane aralık modu vardır. Standart çekim modunda 800 mm ve 4000 mm, yakın çekim modunda ise 400 mm ve 3000 mm arasındaki noktaları tespit edilebilmektedir. Buna ek olarak, aralık dışı değerler SDK ile özel değerlere geri döndürülebilmektedir. 400 mm'nin altındaki ve 8000 mm'nin üzerindeki değerler *bilinmeyen* olarak işaretlenir. Yakın çekim modunda, 3000 mm ve 8000 mm arasındaki değerler *çok uzak* olarak işaretlenirken, standart çekim modunda 4000

mm ve 8000 mm arasındaki değerler *çok uzak*, 400 mm ve 800 mm arasındaki değerler *çok yakın* olarak adlandırılır [32].

Derinlik akışı (stream) 16-bit değerlerini kullanarak verileri depolar. Her pikselin 13 yüksek-derecede bitleri, kamera düzlemi ve en yakın nesne arasındaki etkili mesafeyi milimetre biriminde içermektedir. Her pikselin 3 düşük-derecede bitleri, geçerli pikselin oyuncu segmentasyon haritasının gösterimini kapsamaktadır [32].

Kinect SDK, yüz üzerinde önceden tanımlanmış noktaların belirlenmesi, bu noktalar kullanılarak ağız şeklinin canlandırmasının gerçekleştirilmesini sağlar. SDK kullanıcı ara yüzleri *(Natural User Interface-NUI)* ile yüz izleme, yüz ifadelerini tanıma gibi uygulamalarda kullanılabilir. Yüz izleme SDK ile yapılabilen uygulamalara örnek olarak şunları gösterebiliriz:

- Bir ya da birden fazla kullanıcıya ait yüz şeklinin tanımlanıp anlık takibinin yapılması.
- Yüz üzerinde önceden tanımlanmış 121 noktanın, anlık olarak üç boyutlu derinlik bilgisini de içerecek şekilde Kinect koordinat uzayında takibinin sağlanması. Yüz tanımlama işleminden elde edilen 121 noktanın görsel örneği Şekil 2.2'de gösterilmektedir.

Şekil 2.2. Kinect SDK ile belirlenen 121 noktanın gösterimi

2.3.K-En Yakın Komşu Algoritması (KNN)

K-En Yakın Komşu Algoritması *(K-Nearest Neighbour-KNN),* durum tabanlı *(instance based)* bir öğrenme algoritması olup, sınıflandırılmamış bir kayıt için sınıflandırma mevcut eğitim setindeki en benzer kayıtları karşılaştırma yöntemine dayanır ve konuşma tanıma sistemlerinde kelimelerin benzerliklerini sınıflandırmak için kullanılır. Algoritma basit olmasına karşın bazı sorunlar içermektedir. Bunlar;

- Komşu sayısı *(k)* için standart bir değer mevcut değildir, deneysel olarak veri setinin yapısı ile değişkenlik gösterebilir. Tahmin hatasını en aza indirgemek için küçük değerlerin verilmesi öngörülmektedir. Bu nedenle başlangıç değeri genellikle 1 olarak kabul edilir.
- Mesafe ölçüm yöntemi belirlenmesi veri setinin yapısı ile ilişkilidir. Kullanılan mesafe sınıflarına örnek olarak; Chebyshev, Öklid, Manhattan ve Minkowski mesafeleri gösterilebilir.

En sık kullanılan uzaklık fonksiyonu olan Öklid *(Euclidean)* uzaklığı gerçek dünyadaki uzaklığı temsil eder. Öklid uzaklığı (2.1)'e göre ifade edilmektedir.

$$d_{\text{Öklid}}(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{2.1}$$

x_i : Karşılaştırılacak ilk kaydın özniteliklerini temsil eder.
y_i : Karşılaştırılacak ikinci kaydın özniteliklerini temsil eder.
n : *x* ve *y* vektörlerinin boyutu.

Kullanılan ikinci uzaklık fonksiyonu Manhattan uzaklığı olup, (2.2)'de gösterilmektedir.

$$d_{Manhattan} = \sum_{i=1}^{n}|x_i - y_i| \tag{2.2}$$

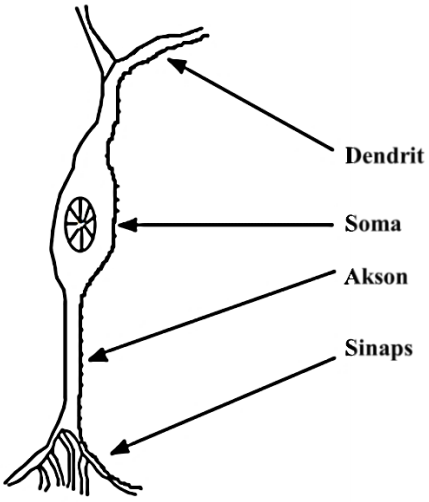
2.4.Yapay Sinir Ağları (ANN)

Yapay Sinir Ağları (ANN), kabaca beynin sinirsel yapısını taklit eden makine öğrenmesi ve örüntü tanıma yeteneğine sahip hesaplama modeli olup, biyolojik sinir sistemini taklit ederler [33]. Biyolojik nöronun genel yapısı Şekil 2.3'te gösterilmektedir.

Yapay sinir ağları, insan beyninin özelliklerini taklit ederek;

- Öğrenme
- İlişkilendirme
- Sınıflandırma
- Genelleme
- Özellik belirleme ve
- Optimizasyon gibi konuları başarılı bir şekilde uygulamaktadır [34].

Yapay sinir ağları biyolojik sinirlerin işlevlerini taklit eder, bir başka deyişle beynin bilgiyi işleme yöntemini taklit eder. Çizelge 2.1'de yapay sinir ağı ile biyolojik sinir ağının işlevlerinin karşılaştırılması gösterilmiştir. İnsan beyni doğumdan itibaren olayları yaşayarak öğrenir. Bu öğrenme tekniği ANN'de de kullanılmaktadır. Öğrenme eğitim yoluyla gerçekleşir, örnekler kullanılarak eğitim verilerinin işlenmesiyle olur [34].



Şekil 2.3. Biyolojik nöronun genel yapısı [35].

Çizelge 2.1. Biyolojik sinir ağı ile yapay sinir ağının karşılaştırması [33]

Biyolojik Sinir Ağı	**Yapay Sinir Ağı**
Sinir Sistemi	Sinirsel Hesaplama Sistemi
Sinir	Düğüm(Sinir, İşlem elemanı)
Sinaps	Sinirler arası bağlantı ağırlıkları
Dendrit	Toplama İşlevi
Hücre Gövdesi	Etkinlik İşlevi
Akson	Sinir Çıkışı

Bir sinir ağının temel işlem elemanı nörondur. Temel olarak bir biyolojik nöron gelen sinyalleri alır, bir şekilde birleştirir, sonuç üzerinde doğrusal olmayan bir işlem gerçekleştirir ve nihai sonucu verir [34].

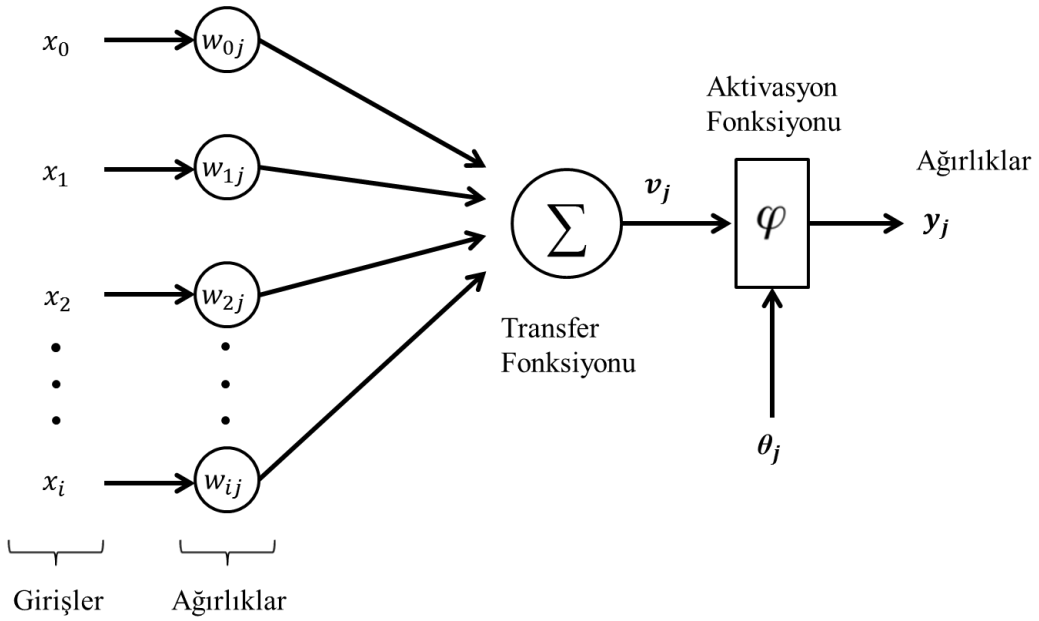
Nöronların temel olarak dört bileşeni bulunmaktadır. Bu bileşenlerin biyolojik isimleri *dendrit*, *akson*, *soma* ve *sinapstır*. Sinapslar, sinir hücreleri arasındaki sinyal akışının geçmesini sağlayan bağlantı elemanlarıdır. Bu sinyaller somaya gider ve burada hücre kendi sinyallerini oluşturarak akson aracılığı ile dendrite gönderir [34].

2.4.1. Yapay sinir ağlarının ana öğeleri

Yapay sinir ağları, insan beyninin gerçekleştirdiği işlevleri simüle etmek için doğal nöronların dört temel işlevini kullanır. Dış ortamdan veya diğer hücrelerden alınan girdiler, ağırlıklar yardımıyla hücreye bağlanır. Toplama fonksiyonu ile net girdi hesaplanır. Net girdinin aktivasyon fonksiyonundan geçirilmesiyle net çıktı hesaplanır. Bu işlem aynı zamanda hücrenin çıkışını verir.

Şekil 2.4'te bir yapay sinir ağının yapısı gösterilmektedir. Giriş sinyalleri x_i sembolüyle ifade edilmiştir. Giriş sinyallerinin her biri w_{ij} ile ifade edilen bağlantı ağırlıkları ile çarpılır. Bu ürünler (v_k), eşik değeri θ_j ile toplanır ve sonuç üretmek için transfer fonksiyonu ile işlendikten sonra y_j çıkışları elde edilir [35].

Girdiler $(x_0, x_1, x_2, \dots, x_i)$, bir yapay sinir hücresine dış dünyadan, başka hücrelerden ya da kendi kendisinden gelen bilgilerdir [34].



Şekil 2.4. Yapay sinir ağı

Ağırlıklar $(w_{0j}, w_{1j}, w_{2j}, \ldots, w_{ij})$, hücreye gelen bilginin önemini ve hücre üzerindeki etkisini gösterir. Bir nöron genellikle eşzamanlı olarak birçok girdi alır. Her girdinin kendisine ait bir ağırlığı vardır. Ağırlık, girdinin hücre üzerindeki etkisini göstermektedir [34].

Toplama fonksiyonu, hücreye gelen her bir girdi değeri kendi ağırlığı ile çarpılarak toplanır ve böylece ağa giren net girdi elde edilmiş olur. Basit şekilde formüle etmek gerekirse;

$$Input_{1j}(x_1) = x_1 * w_{1j} \quad (2.3)$$

$$Input_{2j}(x_2) = x_2 * w_{2j} \quad (2.4)$$

$$Input_i(x_i) = x_i * w_{ij} \quad (2.5)$$

$$\vdots \qquad \vdots$$

$$toplamInput = \sum_i^n x_i w_{ij} \quad (2.6)$$

Yapay sinir ağlarında daima bu formül kullanılmaz. Yapay sinir ağı modeline göre basit çarpım toplamı işlemi yerine toplama fonksiyonu maksimum, minimum, çoğunluk ve kümülatif toplamı da kullanabilir [34].

Aktivasyon fonksiyonu, toplama fonksiyonun sonucu olarak hücreye gelen net girdiyi işleyerek çıktıyı belirler. Bir problem için en uygun fonksiyon tasarımcının denemeleri sonucunda belirlenebilir, uygun fonksiyonu elde etmek için bir formül kullanılmamaktadır. Yaygın olarak sigmoid fonksiyonu kullanılmaktadır. Bunun dışında lineer fonksiyon, step fonksiyonu, sinüs fonksiyonu, eşik değer fonksiyonu ve hiperbolik tanjant fonksiyonu da kullanılmaktadır. Şekil 2.5'te örnek aktivasyon fonksiyonları gösterilmiştir [34].

Sigmoid aktivasyon fonksiyonu, türevi alınabilir ve sürekli bir fonksiyondur. Giriş değeri toplama fonksiyonu kullanılarak belirlenmektedir. Giriş değerlerinin her biri için sıfır ile bir arasında bir değer üretir:

$$f(x) = \frac{1}{1 + e^{-x}} \tag{2.7}$$

Hiperbolik tanjant aktivasyon fonksiyonu, giriş değerleri tanjant fonksiyonundan geçirilerek hesaplanır. Fonksiyonun çıkış değerleri -1 ile 1 arasındadır.

$$\tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \tag{2.8}$$

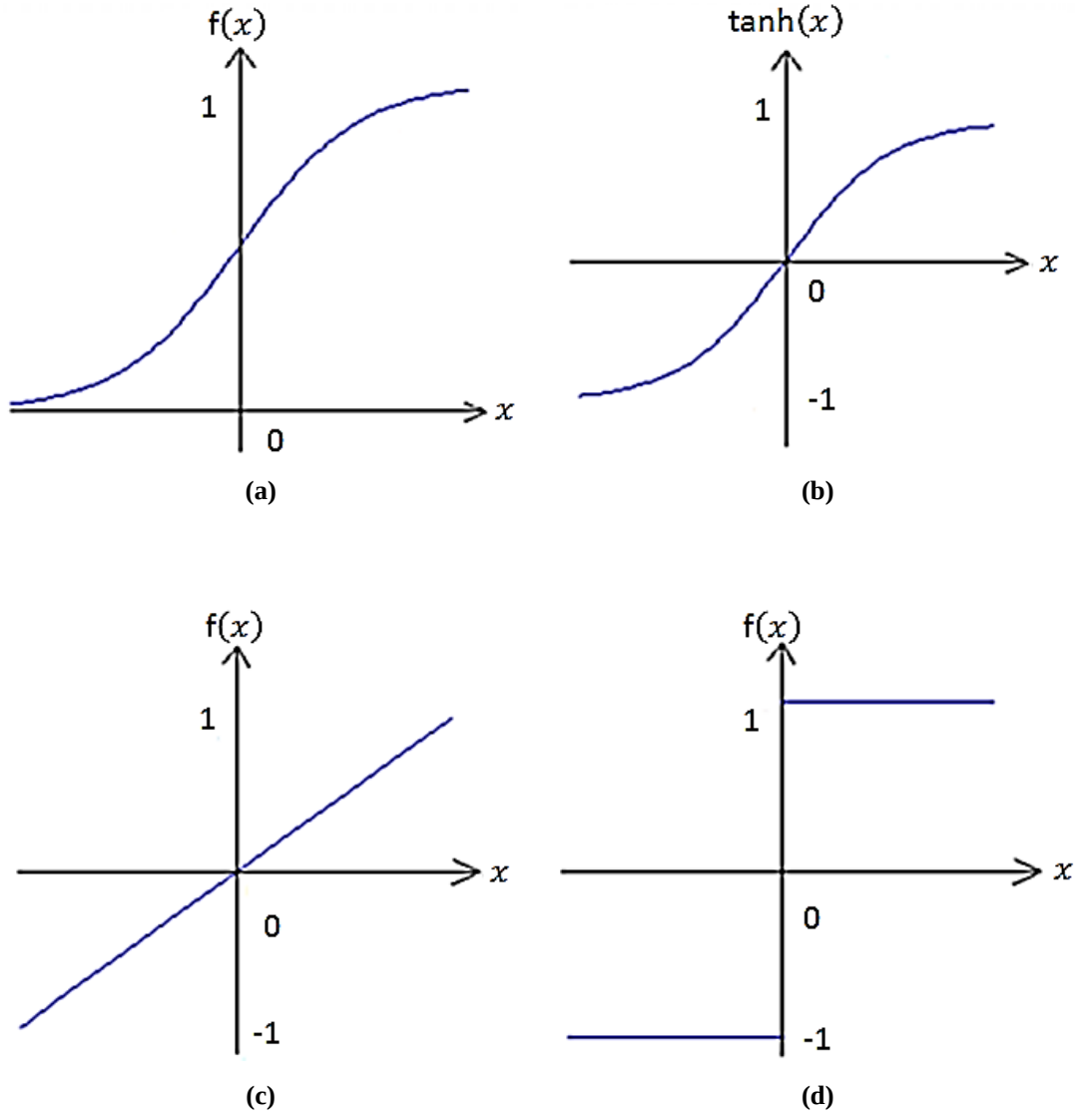
Doğrusal aktivasyon fonksiyonu, doğrusal problemleri çözmek için kullanılır. Gelen girdiler hücrenin çıktısı olarak kabul edilir:

$$f(x) = A.x \tag{2.9}$$

Adım aktivasyon fonksiyonu, gelen girdilerin belirlenen eşik değerinden büyük ya da küçük olmasına göre 1 ya da 0 olarak değer alır:

$$f(x) = \begin{cases} -1, & x \leq 0 \\ 1, & x > 0 \end{cases} \tag{2.10}$$

Hücrenin çıktısı, aktivasyon fonksiyonu tarafından belirlenen dış dünyaya ya da diğer sinirlere gönderilen çıktı değeridir. Her bir çıkışta birçok nörona çıktı olarak gönderilebilecek tek bir çıktı üretebilir. Bu yapı biyolojik nöronlar ile aynı özellikleri taşımaktadır, yani biyolojik sinirde olduğu gibi birçok giriş varken sadece tek bir çıkış etkinliği bulunmaktadır [35].



Şekil 2.5. Aktivasyon fonksiyonu örnekleri. a) Sigmoid aktivasyon fonksiyonu, b) Hiperbolik tanjant aktivasyon fonksiyonu, c) Doğrusal aktivasyon fonksiyonu ve d) Adım aktivasyon fonksiyonu

2.4.2. Yapay sinir ağlarında öğrenme

Yapay sinir ağlarında bilgi, ağdaki bağlantı ağırlıklarında depolanır ve öğrenme işlemi ağdaki bu ağırlıkların hesaplanması ile istenilen işlemi gerçekleştirebilir. Sinirler üzerindeki bağlantı ağırlıkları, belirlenen öğrenme kuralları aracılığı ile bulunarak ağ eğitilmiş olur.

Öğrenme yöntemleri temelde *danışmanlı* (supervised) ve *danışmansız* (unsupervised) olarak iki başlık altında toplanabilir. Çizelge 2.2'de danışmanlı ve danışmansız öğrenme yöntemleri için örnekler gösterilmiştir.

Danışmanlı öğrenme, gerçek çıkış ile istenilen çıkış arasında daha yakın karşılaştırmalar üretmek için ağırlıkların öğrenme yöntemleri kullanılarak tanımlanmasıdır. Öğrenme yönteminin amacı anlık hatayı en aza indirmektir. Bu işlemi gerçekleştirmek için belirlenen ya da kabul edilebilir hataya ulaşana kadar bu işlem tekrarlanır. Eğitim işlemi her giriş kümesi için uygun çıkış kümesi ağa öğretilerek sağlanır. Her örnek için giriş ve çıkış işlemleri sisteme öğretilir.

Danışmansız öğrenmede ise, sistemin doğru çıkış hakkında bilgisi yoktur. Ağa verilen girişlere göre çıkış işlemini kendisi üretir. Danışmansız öğrenmede ağ girilen giriş verileri ile çalışır. Örneklerdeki parametreler arasındaki bağlantıları ağın kendi kendisine öğrenmesi beklenmektedir [33].

Çizelge 2.2. Danışmanlı-danışmansız öğrenme yöntemleri [33]

Danışmanlı Öğrenme	**Danışmansız Öğrenme**
Perceptron	Kümeleme (Örneğin, K-Means kümeleme)
Delta Öğrenme	Saklı Markov Modeli
Geri Yayılım	Temel Bileşen Analizi
Çok Katmanlı Perceptron	Hebbian Öğrenme

2.4.3. Geri yayılım algoritması

Geri Yayılım *(Back Propagation)* günümüz yapay sinir ağlarında yaygın olarak kullanılan bir öğrenme algoritmasıdır. Tipik bir geri yayılım ağı bir giriş, bir çıkış ve en az bir gizli katmandan meydana gelmektedir. Şekil 2.6'da üç

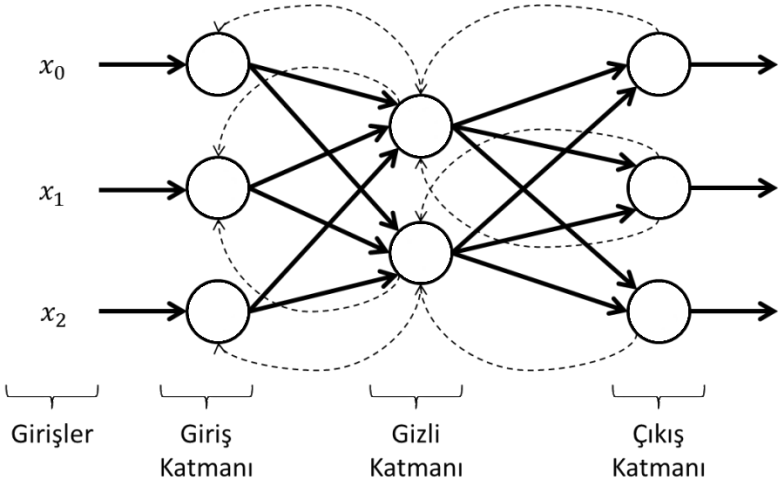
düğümlü giriş katmanı, iki düğümlü bir gizli katmandan ve üç düğümlü çıkış katmanından meydana gelen bir geri yayılım ağı örneği gösterilmektedir. Gizli katman sayısı problemin yapısı ile ilgili değişkenlik göstermektedir. Ancak bazı çalışmalar göstermiştir ki, problemleri çözmek için üç gizli katman ve bir de çıktı katmanını içeren en az dört katman bulunması gerekmektedir [33].

Katman sayıları ve her katmandaki düğüm sayıları ağın başarım ölçütünde önemli bir role sahiptir. Belirli bir uygulama için ağın yapısı ile ilgili ölçülebilir bir optimum cevap yoktur. Zaman içerisinde araştırmacıların mevcut sorunlar için referans aldıkları genel kurallar bulunmaktadır.

Kural 1: Giriş ve çıkış verileri arasındaki ilişkinin karmaşıklığı ile doğru orantılı olarak gizli katmandaki işlem elemanlarının da sayısı arttırılmalıdır [33].

Kural 2: Eğer modellenen işlem birden fazla aşamaya ayrılabiliyorsa gerekli olan gizli katman sayısı arttırılmalıdır. Bunun aksine fazla sayıda kullanılan gizli katman sayısı modellenen işlem aşamalara ayrılamıyorsa sistemde ezberlemelere yol açar ve yanlış sonuçlara neden olur [33].

Kural 3: Gizli katmandaki işleme sayının belirlenmesi için, eğitim verisinin miktarı önemli bir parametredir. Üst sınırı belirlemek için eğitim kümesindeki girdi-çıktı çiftlerinin sayısı ağdaki toplam giriş ve çıkış düğüm sayısına bölünür [33].



Şekil 2.6. Geri yayılım ağı örneği

Q katmanlı, ileri beslemeli bir ağ için geri yayılım algoritması aşağıda anlatılmıştır. Kullanılan değişkenler;

$q = 1,2,\dots,Q$: Katman numarası

H_i^p : q'uncu katmandaki *i* biriminin girdisi

y_i^q : q'uncu katmandaki *i* biriminin çıktısı

$w_{ij}^q = (q-1)$: *q-1*'inci katmandaki *i* birimini q'uncu katmandaki *j* birimine bağlayan ağırlık

η : Öğrenme katsayısı

1. *w* için başlangıç değerleri atanır.

2. Rastgele bir $\{x^p, t^p\}$(giriş-çıkış) çalışma modeli seçilip ve seçilen bu model kullanılarak q katmanındaki her bir *j* için ileri yönde çıktı hesaplanır. Bu işlem sonrasında çıkış;

$$y_i^q = f\left(\sum_i y_i^{q-1} w_{ij}^q\right) \tag{2.11}$$

$$y_i^0 = x_i \tag{2.12}$$

3. Çıkış birimleri için hata terimleri hesaplanır.

$$\sigma_i^Q = \left(v_i^Q - y_i^p\right) f'\left(H_i^Q\right) \tag{2.13}$$

4. Geriye yayılımla hata terimleri hesaplanır.

$$\sigma_i^{q-1} = f'\left(H_i^{q-1}\right) \sum_k \sigma_i^q w_{ij}^q \tag{2.14}$$

5. Bütün ağırlıklar $\boldsymbol{w_{ij}}$ kullanılarak güncellenir.

$$w_{ij}^{yeni} = w_{ij}^{eski} + \Delta w_{ij}^q \tag{2.15}$$

$$\Delta w_{ij}^q = \eta \sigma_i^q y_i^{q-1} \tag{2.16}$$

6. Toplam hata kabul edilebilir düzeye gelene kadar 2. adıma dönülerek her bir *p* modeli için işlem tekrarlanır [33].

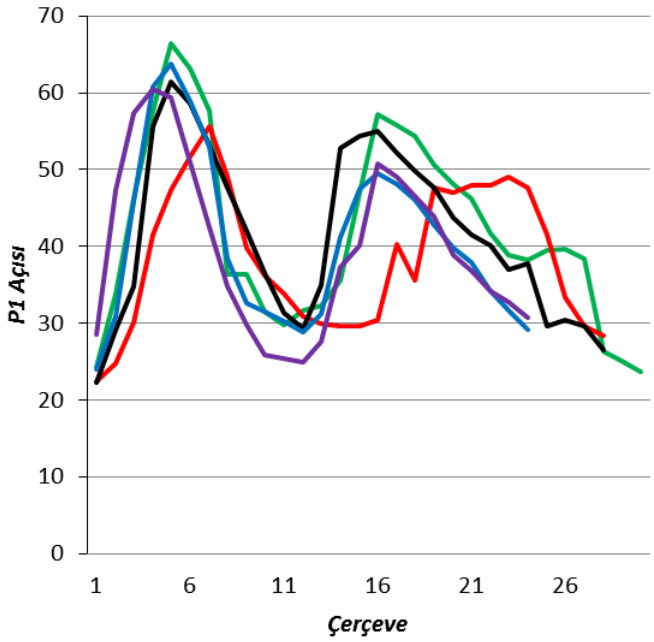
2.5.Dinamik Zaman Bükmesi

Dinamik Zaman Bükmesi *(Dynamic Time Warping-DTW)* belirli kısıtlamalar altında verilen zamana bağlı iki dizi arasındaki optimum benzerliği bulmak ve özellikle ses sinyallerinde iki sinyalin birbirine olan benzerliğini bulmak için kullanılan bir yöntemdir [15]. DTW otomatik konuşma tanıma sistemlerinde farklı konuşma kalıplarını karşılaştırmak için kullanılır. Bu diziler, ayrık sinyaller, zaman serileri ya da zaman içerisinde eşit uzaklıklarla örneklenmiş öznitelik dizileri olabilir [36].

Uygulamada konuşulan kelimeler izole edilmiş kelimelerdir, yani konuşma metinleri arasında konuşmanın başlangıç ve bitiş noktaları arasında yeterince sessizlik bulunmaktadır. Bu türde tanımlanmış kelimelerin başlangıç ve bitiş noktalarının tahminleri oldukça kolaydır. Sürekli Konuşma Tanıma *(Continuous Speech Recognition)* sistemlerinde, kişilerin konuşması günlük yani doğal konuşma şekliyle aralıksız olarak gerçekleştiği için kelime başlangıç ve bitiş noktasını belirlemek kolay olmayabilir. Söylenen kelimeler tek kişi tarafından söyleniyor ve sistem bu konuşmacıya ait verilerle sınıflandırma yapıyorsa buna *kişiye bağlı sistem* ismi verilir. Bundan daha karmaşık olarak nitelendirilebilecek sistem *kişiden bağımsız sistem*dir. Kişiden bağımsız olarak çalışan sistemlerde, sistem bir dizi farklı konuşmacı tarafından eğitilmeli ve eğitim için kullanılan konuşmacılardan farklı kişiler tarafından sağlanan verilerin sınıflandırmasını yapabilmesi beklenmektedir [37].

Herhangi bir İzole Edilmiş Kelime Sınıflandırma *(Isolated Word Recognition-IWR)* sisteminin temelinde bir dizi önceden tanımlanmış referans örüntü ve bu örüntülere ait uzaklık ölçüleri yer almaktadır [37].

Bir kullanıcının "*beyaz*" kelimesini 5 kez ardı ardına tekrar ettiğinde alt ve üst dudağın hareketini ifade eden, açılıp kapanma açısının kelime telaffuzu süresince değişimi Şekil 2.7'de gösterilmektedir. Kullanıcının ve söylenen kelimelerin aynı olmasına rağmen; kelimenin başlangıç ve bitiş noktaları, durağanlıklar, kelime telaffuz uzunlukları ve dudağın açılış kapanış açıları birbirinden farklılık göstermektedir.



Şekil 2.7. Beyaz kelimesinin aynı kullanıcı tarafından 5 kez tekrarı

Bunun basit bir doğrusal zaman ölçekleme olmadığını söylemek gerekmektedir. Söylenen bu kelimelerin aynı kişi tarafından söylenen aynı kelimeler olduğunu eşleştirmek için doğrusal olmayan haritalama *(highly nonlinear mapping)* gereklidir.

Referans ve test örüntüleri optimal şekilde eşleştirmek için gerekli olan doğrusal olmayan haritalamayı çözümlemek için dinamik programlama tekniklerine başvurulmuştur.

2.5.1. DTW kısıtlamaları

Sınırlandırma (Boundary): Zaman serileri X ve Y'nin ilk ve son elemanları birbirleriyle eşleşmelidir. Eğrilme yolu $w_k = (X, Y)$ yani bu zaman serilerinin uzunlukları olan $|X|$ ve $|Y|$' den başlar $w_1 = (1,1)$'e kadar devam eder.

Süreklilik (Continuity): X ve Y'deki bütün elemanlar kullanılır ve aynı zamanda hizalamada tekrar olmayacağını gösterir

Monotonluk (Monotonicity): Eğrilme fonksiyonu artan monoton bir fonksiyon olmak zorundadır.

2.5.2. Problemin formülasyonu

Verilen iki zaman serisi *X* ve *Y*, bu zaman serilerinin uzunlukları $|X|$ ve $|Y|$ olarak tanımlanmıştır.

$$X = x_1, x_2, .., x_i, .., x_{|X|} \tag{2.17}$$

$$Y = y_1, y_2, .., y_j, .., y_{|Y|} \tag{2.18}$$

Optimum eğrilme yolu *W* hesaplanmıştır.

$$W = w_1, w_2, .., w_k \tag{2.19}$$

$$\max(|X|, |Y|) \leq K < |X| + |Y| \tag{2.20}$$

K: Optimum eğrilme yolu uzunluğu.

w_k: Optimum eğrilme yolunun k'ıncı elemanı.

w: Kılavuz noktalar dizisidir, her w_k bir (i, j) noktasına karşılık gelir.

$$w_k = (i, j) \tag{2.21}$$

i : Zaman serisi X'in indeksi.

j : Zaman serisi Y'nin indeksi.

Optimum eğrilme yolu her bir zaman serisinin başından başlamalıdır ($w_1 = (1,1)$) ve zaman serilerinin sonunda bitmelidir ($w_k = (|X|, |Y|)$). Bu gösterim, her iki zaman serisinin indeksinin eğrilme yolunun kullanıldığını garanti altına alır.

Optimum eğrilme yolu için bir sınırlama bulunmaktadır. Bu kısıtlama *i* ve *j*'yi monotonik olarak eğrilme yolunda arttırılmaya zorlar. Her zaman serisinin her indeksi kullanılmalıdır.

$$w_k = (i, j), w_{k+1} = (i', j') \tag{2.22}$$

$$i \leq i' \leq i + 1, \quad j \leq j' \leq j + 1 \tag{2.23}$$

Optimum eğrilme yolu minimum uzaklığa sahip eğrilme yoludur ve Dist(W) ile gösterilir.

$$\text{Dist}(w) = \sum_{k=1}^{k=K} \text{Dist}(w_{ki}, w_{kj}) \tag{2.24}$$

$\text{Dist}(w)$: W eğrilme yolunun uzaklık

$\text{Dist}(w_{ki}, w_{kj})$: Eğrilme yolunun k'ıncı elemanı için iki nokta arasındaki uzaklık [38,39].

2.5.3. DTW Algoritması

Minimum uzaklık eğrilme yolunu bulmak için dinamik programlama yaklaşımları kullanılmaktadır. Bütün problemi bir seferde çözmek yerine problem alt-problemlere ayrılır.

$X' = x_1, x_2, \ldots, x_i$ ve $Y' = y_1, y_2, \ldots, y_j$ zaman serilerinden elde edilen minimum uzaklık eğrilme yolu olan D(i, j) değeri elde edildiğinde, iki $|X|$ ve $|Y|$ birikmiş maliyet matrisi *D* elde edilmektedir.

$$D(i, j) = \text{Dist}(i, j) + \min[D(i-1, j), D(i, j-1), D(i-1, j-1)] \tag{2.25}$$

Son olarak birikmiş maliyet matrisi elde edildikten sonra optimum eğrilme yolu elde edilir [38,39].

$$W = w_1, w_2, .., w_k \tag{2.26}$$

$$max(|X|, |Y|) \leq K < |X| + |Y| \tag{2.27}$$

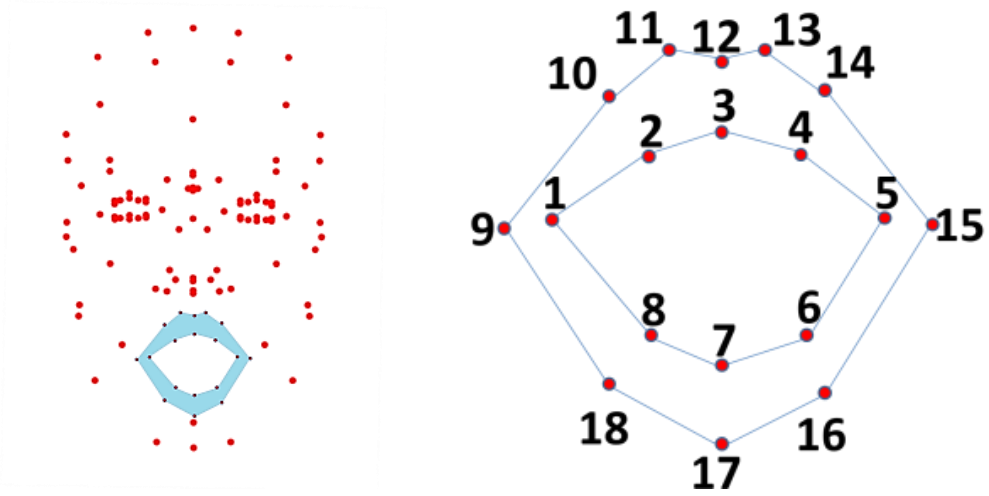
$$w_{k-1} = \begin{cases} (1, j-1), & i == 1 \\ (i-1, 1), & j == 1 \\ \text{argmin}\{D(i-1, j-1), D(i-1, j), D(i, j-1)\}, & \text{diğer durumlar} \end{cases} \tag{2.28}$$

3. YAPILAN ÇALIŞMALAR ve VERİ SETİNİN OLUŞTURULMASI

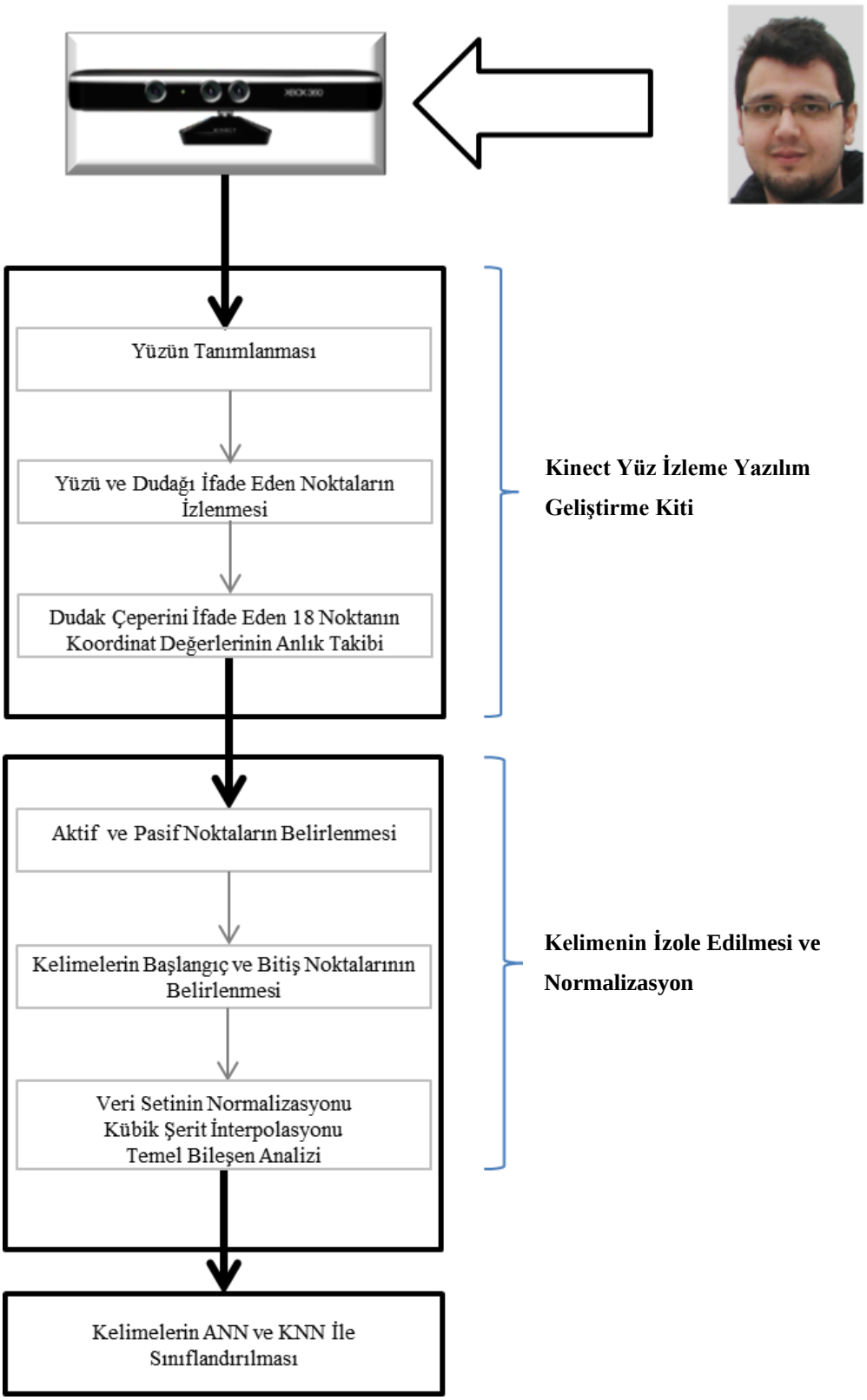
Bu çalışmada, herhangi bir eğitim verisi kullanılmadan söylenen iki kelimenin birbirine olan benzerliğini karşılaştırmak için kişiden bağımsız olarak kullanılabilecek parametreleri bulmak hedeflenmiştir. Bu nedenle, öncelikli olarak kişinin kamera karşısındaki konuşmasını kayıt altına alıp, elde edilen görüntüler üzerinde belirlenen noktaların takibini yapacak bir sisteme ihtiyaç duyulmaktadır.

Kinect kamerası ve Kinect SDK 1.7 versiyonu bu işlemleri gerçekleştirmek için kullanılmıştır. Bu çalışmada, Kinect Face Tracking SDK kullanılıp Kinect kamerasından gelen görüntülerin analizi yapılarak Şekil 3.1'de gösterilen yüz üzerinde önceden belirlenmiş dudağı ifade eden 18 noktanın anlık takibini kaydeden bir yazılım geliştirilmiştir.

Bu veriler kullanılarak telaffuz edilen kelimelere ait bir veri seti oluşturulmuştur. Elde edilen veri seti ile K-en yakın komşu yöntemi ve geri yayılımlı yapay sinir ağı kullanılarak kelimelerin sınıflandırılma işlemi yapılmıştır. Böylece kayıt işlemi ve veri toplama işleminin başarısı test edilmiştir. Yapılan çalışmanın genel yapısı Şekil 3.2'de gösterilmiştir.



Şekil 3.1. a) Yüz üzerinde tanımlanmış 121 nokta ve b) Dudağı ifade eden 18 nokta



Şekil 3.2. Görsel verilerin elde edilmesi ve sınıflandırma işlemi

Sınıflandırma başarıları doğrultusunda, KNN algoritması ile kapsamlı arama yöntemi *(exhaustive search method)* kullanılarak her kullanıcı için ayırt edici ve sınıflandırma başarısı yüksek özniteliklerin seçimi yapılmıştır. Ayrıca geri yayılımlı yapay sinir ağı modeli ile elde edilen veri setinin tutarlılığı test edilmiştir. Yapılan testler ve elde edilen sonuçlar doğrultusunda kelime benzerliklerini hesaplayabilecek bir sistem önerilmiştir. Önerilen bu sistemde, dinamik zaman bükmesi algoritması kullanılarak ve ardı ardına söylenen iki kelimenin aynı olduğu varsayımı ile yola çıkarak iki kelimenin birbirleri ile benzerlik oranlarını hesaplanmıştır.

3.1.Görsel Verilerin Elde Edilmesi

Veri setini oluşturmak için Türkçe'de sıklıkla kullanılan renk isimleri seçilmiştir. Seçilen bu renkler sırasıyla; beyaz, bordo, gri, kahverengi, kırmızı, lacivert, mavi, menekşe, mor, pembe, sarı, siyah, turkuaz, turuncu ve yeşildir. Veri toplama işlemi sırasında her bir kelime aynı kullanıcı tarafından ardı ardına beşer kez tekrarlanmıştır. Bu tekrarlamalar arasında bir kelimenin söylenmeye başlanıp bitirilmesi istenen süre 4 saniye olarak belirlenmiştir. Söylenecek kelime ve kelimenin söylenme anı görsel ve işitsel uyarılar ile kullanıcıya belirtilmiştir.

Konuşma esnasında dudağın anlık olarak hareketlerini yakalamak için Kinect Face Tracking Engine kullanılmıştır. Görsel verileri kayıt etmek için Kinect Face Tracking Basics yazılımı temel alınmıştır. Gerçek zamanlı veri kayıt işlemi için Microsoft Visual Studio .Net platformunda C# yazılım dili kullanılarak sistem geliştirilmiştir. Yüz yakalama sistemi tarafından önceden tanımlanmış 121 noktanın Kinect koordinat düzlemindeki sayısal koordinat değerleri *(x,y,z)* anlık olarak kayıt altına alınmıştır. Bu noktalar içerisinden Şekil 3.1(b)'de gösterilen 18 nokta belirlenmiştir. Belirlen bu noktalardan 8 tanesi dudağın iç çeperini, diğer 10 tanesi ise dudağın dış çeperini ifade etmek için kullanılmıştır. Dudağın iç çeperini ifade eden Şekil 3.1(b)'de gösterilen temsili referans numaraları; 1,2,3,4,5,6,7 ve 8'dir. Dudağın dış çeperini ifade eden Şekil 3.1b'de gösterilen temsili referans numaraları ise; 9,10,11,12,13,14,15,16,17 ve 18'dir.

Kameradan elde edilen görüntüler 640x480 piksel çözünürlüktedir. Bu görüntüler 30 fps kalitesi ile kayıt altına alınmıştır. Ayrıca belirlenen noktaların derinlik bilgilerini de içerecek şekilde 3 boyutlu Kinect koordinat düzlemindeki sayısal koordinat değerleri de kaydedilmiştir. Dudak hareketlerini tanımlamak için, noktaların koordinat düzlemi üzerindeki değerleri kullanılarak, bu noktalardan oluşan vektörlerin açı değerleri oluşturulmuştur.

Kayıt öncesinde kullanıcılar başlangıç için standart hale getirilmiş bir uzaklık ve kafa pozisyonu ile kamera karşısına konumlandırılmışlardır. Fakat konuşma esnasında istem dışı anlık hareket bozuklukları meydana gelmektedir. Bu hareketlerden meydana gelen tutarsızlıkları en aza indirmek için, noktaların anlık değerleri yerine bu noktalardan oluşan vektörler arasında meydana gelen açı değerleri kullanılmıştır [23]. Ham koordinat verisini kullanmak yerine açı değerleri kullanmanın nedenleri aşağıdaki şekilde sıralanabilir:

- Bir kullanıcıdan veri toplamak için kamera karşısındaki kayıt süresi yaklaşık olarak beş dakikadır. Bu süre içerisinde kullanıcının kamera karşısındaki kafa pozisyonunu hiç değiştirmeden koruyabilmesi mümkün olmamaktadır. Bu nedenle konuşma esnasında kullanıcı kafa pozisyonu ile kamera arasındaki açı değişimlere uğramaktadır. Bu değişimler, toplanan veride tutarsızlıklara neden olmaktadır.
- Başlangıç için standart hale getirilen kullanıcı-kamera uzaklığı, kayıt süresince değişimlere uğramaktadır. Örneğin bir kullanıcı kameraya 50 cm'lik bir uzaklığa oturup kayıt yapıldığında toplanan veriler ile aynı kullanıcının kameraya 60 cm uzaklıkta oturduğu zaman toplanan veriler, kinect koordinat düzlemi için kayıt yapıldığında sayısal olarak farklılıklar gösterecektir.
- Şekil 3.3'te bir kullanıcının kamera karşısındaki görüntüsü gösterilmektedir. Kinect koordinat uzayı orijin noktası olarak sensör ile RGB kameranın orta noktasını kabul eder. Bu uzaya göre kişi kamera ile *z* düzlemine paralel şekilde konumlandırılmalıdır. Bunun aksi bir durumda aynı kullanıcı aynı kelimeyi tekrar ederken kamera orijininden yapılan sapmalar sayısal koordinat değerlerini değiştireceği için veri setinde tutarsızlıklar meydana gelecektir.

- Ayrıca konuşma esnasında kullanıcıların fiziksel farklılıkları da sayısal olarak farklı değerlere neden olmaktadır. Fiziksel olarak dudağın boyutları ve konuşma esnasında kullanıcının günlük konuşma alışkanlıkları ile değişkenlik gösteren alt ve üst çeperin baskın olarak istenilenden fazla açılıp kapanması yine sayısal olarak farklılıklara neden olacaktır.

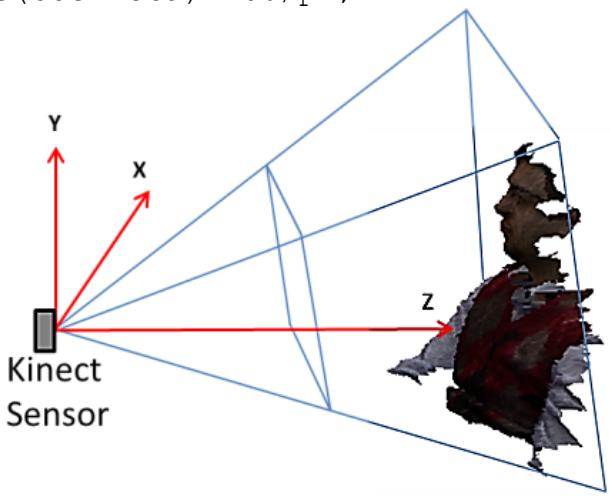
Yukarıda tanımlanan problemleri en aza indirmek için kamera görüntülerinden elde edilen noktaların Kinect koordinat uzayındaki sayısal değerleri, belirlenen noktalar ile elde edilen vektörlerin açı değerlerine dönüştürülmüştür. Bu noktalardan elde edilen vektörler ve bu vektörleri kullanarak elde edilen açılar Çizelge 3.1'de gösterilmektedir.

Vektörler arasında kalan açıları hesaplamak için v ve w vektörlerinin iç çarpımı $(v|w)$, vektörlerin normlarının çarpımına bölünmüş ve sonucunda ters kosinüsü alınmıştır.

$$\theta = \arccos\left(\frac{(v|w)}{||v||\,||w||}\right) \tag{3.1}$$

$(x_1, y_1, z_1), (x_2, y_2, z_2)$ ve (x_3, y_3, z_3) noktaları arasında kalan açı derece cinsinden aşağıdaki MATLAB kodu ile hesaplanabilir:

```
v = [x1-x2,y1-y2,z1-z2];
w = [x3-x2,y3-y2,z3-z2];
cosTheta = dot(v,w)/(norm(v)*norm(w));
theta= acos(CosTheta)*180/pi;
```



Şekil 3.3. Kinect koordinat uzayı ve bir kullanıcının kamera karşısındaki derinlik verisi kullanılarak oluşturulan görüntüsü

Çizelge 3.1. Vektörlerin açısal kombinasyonları

Sıra No	Kullanılan Noktalar	v vektörü	w vektörü
1	12-9-17	$v = (9,12)$	$w = (9,17)$
2	12-17-9	$v = (17,12)$	$w = (17,9)$
3	17-12-9	$v = (12,17)$	$w = (12,9)$
4	11-1-17	$v = (1,11)$	$w = (1,17)$
5	1-11-17	$v = (11,1)$	$w = (11,17)$
6	1-17-11	$v = (17,1)$	$w = (17,11)$
7	12-9-15	$v = (9,12)$	$w = (9,15)$
8	9-12-15	$v = (12,9)$	$w = (12,15)$
9	17-9-15	$v = (9,17)$	$w = (9,15)$
10	9-17-15	$v = (17,9)$	$w = (17,15)$
11	3-1-2	$v = (1,3)$	$w = (1,2)$
12	7-3-1	$v = (3,7)$	$w = (3,1)$
13	1-7-3	$v = (7,1)$	$w = (7,3)$
14	1-12-5	$v = (12,1)$	$w = (12,5)$
15	5-1-12	$v = (1,5)$	$w = (1,12)$
16	17-1-5	$v = (1,17)$	$w = (1,5)$
17	1-17-5	$v = (17,1)$	$w = (17,5)$
18	3-1-5	$v = (1,3)$	$w = (1,5)$
19	1-3-5	$v = (3,1)$	$w = (3,5)$
20	17-16-18	$v = (16,17)$	$w = (16,18)$
21	18-17-16	$v = (17,18)$	$w = (17,16)$

3.2.Kelimenin İzole Edilmesi ve Veri Setinin Oluşturulması

Veri toplama işleminde kullanıcının bir kelimeyi söylemeye başlaması ve bitirmesi için tasarlanan zaman 4 saniyedir. Yapılan çalışmalar sırasında kullanıcının bir kelimeyi söylemeye başlaması ve bitirmesi arasında geçen ortalama süre 0,75 saniyedir. Saniyede 30 çerçeve ile görüntü kayıt edildiği için bir kelime yaklaşık olarak 22 çerçeveden meydana gelmektedir. Yani bir kelimenin telaffuzu 1 saniyeden daha az sürmektedir. 4 saniyelik iki sinyal arasında kalan diğer çerçeveler, beklemelerden yani veri setinde kullanılmayacak gürültülerden meydana gelmektedir.

Kelimenin tam olarak başlangıç ve bitiş noktasını bulmak konuşma tanıma sistemlerinin öncelikli problemlerindendir. Tutarlı olarak bu aralığı belirlemek sınıflandırma başarısında önemli bir etkendir. Yapılan çalışmada, kelimenin söylendiği aralığı belirlemek için kaydedilen her çerçeve için anlık enerji ve standart sapma hesapları yapılmıştır. Yapılan işlem sonrasında elde edilen değerler kullanılarak kelimenin telaffuz edildiği aktif konuşma aralığı belirlenmiştir. Bu işlem sonrasında kullanıcıların telaffuz sürelerinden ve konuşma biçimlerinden kaynaklanan değişiklikler normalize edilerek veri seti oluşturulmuştur.

3.2.1. Anlık enerji ve standart sapma ile aktif-pasif noktaların belirlenmesi

Yapılan çalışmada, kelimenin başlangıç ve bitiş noktalarını belirmek için anlık enerji özniteliği ve noktalar arasındaki standart sapma değerleri kullanılmıştır. Enerji özniteliği kelimenin başlangıç-bitiş noktalarının bulmasında ve konuşma süresince sessiz kalınan noktaların belirlenmesinde kullanılan ayırt edici bir özniteliktir [37].

Bu çalışmada, konuşma sırasında dudak çeperinde en fazla deformasyon iki dudak arasındaki açılma-kapanma yani yüksekliği ile genişliğinde meydana gelmektedir. Şekil 3.4'te *P*1 ve *P*2 ile ifade edilen açılar hareketin en baskın şekilde oluştuğu açılardır.

*P*1 ve *P*2 özniteliği için, *N* çerçeveden oluşan bir **X** dizisinin *i*'inci elemanının enerjisi *E(i),* (3.2) ile hesaplanır. Hesaplanan enerji değerleri deneysel olarak belirlenen eşik değeri 0,9 ile karşılaştırılmıştır. Başlangıç için belirlenen enerji eşik değerleri her kullanıcı için farklılık göstermektedir. Bu nedenle, belirlenen eşik değeri hassasiyeti arttırmak için değişkenlik göstermektedir.

Eğer i'inci çerçevenin *E(i)* enerji değeri belirlenen eşik değerinden büyük ise aktif (1), diğer durumlarda ise pasif (0) olarak işaretlenmiştir.

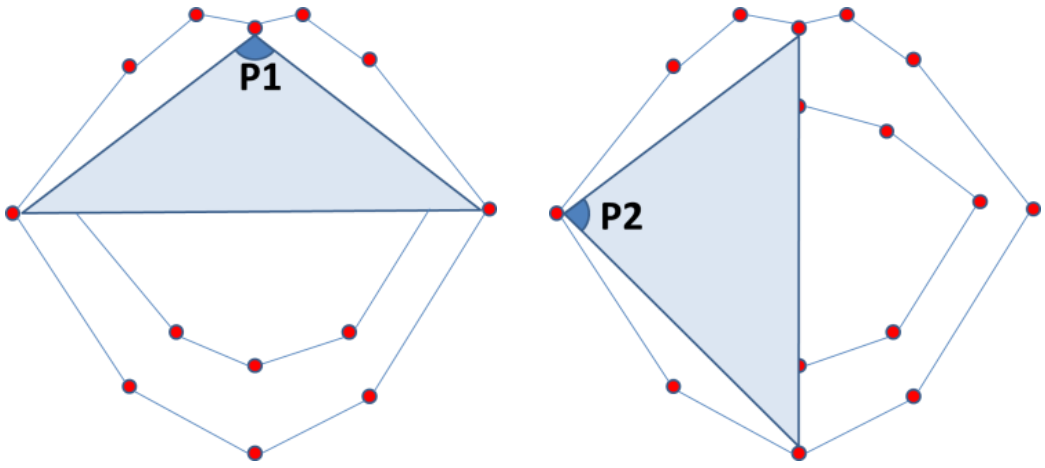
$$\mathrm{X} = (x_1, x_2, .., x_i, .., x_{|X|}) \tag{3.1}$$

$$E(i) = \frac{1}{2k+1} \sum_{j=i-k}^{i+k} |x_i(k)|^2 \tag{3.2}$$

Yapılan çalışmada enerji hesaplama işlemi ile birlikte bir noktanın belirlenen zaman aralığı içerisindeki standart sapması da kullanılmıştır. Noktanın standart sapmasını belirlemek ve noktanın o anda aktif ya da pasif olduğunu anlayabilmek için, kendisinden önceki ve sonraki deneysel olarak belirlenmiş *k* görüntü çerçevesini kullanarak standart sapması hesaplanmıştır. *P*1 ve *P*2 özniteliğinin i'inci çerçevedeki standart sapmasını $\sigma(i)$ hesaplamak için, (3.3) kullanılmıştır. Burada $\overline{p_1}$ belirlenen aralıktaki ortalama değeri ifade etmektedir. Kullanılan *k* değeri deneysel olarak 4 seçilmiştir.

$$\sigma(i) = \sqrt{\frac{1}{2k+1}\sum_{j=i-k}^{i+k}[p_1(j) - \overline{p_1}]^2} \tag{3.3}$$

Bu işlem sonrasında noktaların anlık olarak sahip oldukları sapmalar elde edilmiş olur. Denklem (3.3)'e göre bütün noktaların belirlenen aralığa göre standart sapması hesaplandıktan sonra deneysel olarak belirlenen ve 1 olarak seçilen eşik değeri ile karşılaştırma işlemi yapılmıştır. Başlangıç için belirlenen ortalama eşik değerleri her kullanıcı için farklılık göstermektedir. Bu nedenle, belirlenen eşik değeri hassasiyeti arttırmak için değişkenlik göstermektedir. Eğer i'inci çerçevenin $\sigma(i)$ standart sapması belirlenen eşik değerinden büyük ise *aktif* (1), diğer durumlarda ise *pasif* (0) olarak işaretlenmiştir.



Şekil 3.4. Enerji ve standart sapma için kullanılan *P1* ve *P2 açısı*

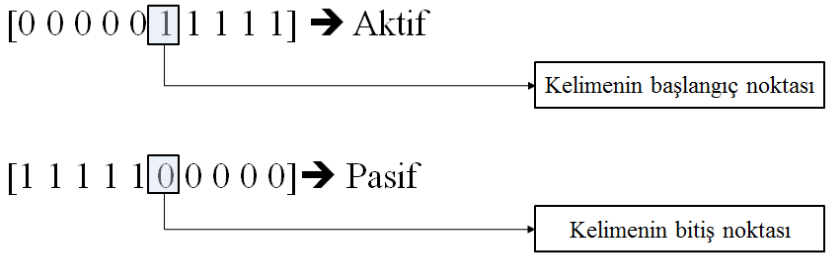
Bir noktanın aktif olarak değerlendirilebilmesi için, enerji ve standart sapma karşılaştırmaları sonrasında elde edilen iki değerin de aktif olması gerekmektedir. Aynı şekilde eğer iki değer de pasif ise bu nokta pasif olarak değerlendirilmektedir.

Son olarak elde edilen enerji ve standart sapma matrisi bir bütün olarak tekrar değerlendirilmektedir. Matristeki bütün aktif ya da pasif noktalar tek tek değerlendirildiğinde; eğer bir pasif noktanın kendisinden 3 çerçeve önceki ve 3 çerçeve sonraki değerleri aktif olarak işaretlenmiş ise, bu nokta aktif olarak kabul edilmektedir.

3.2.2. Kelimelerin başlangıç ve bitiş noktalarının işaretlenmesi

Noktaların aktif ve pasif olarak işaretlenme işlemi sonrasında, elde edilen veriler üzerinde kelimenin başlangıç ve bitiş noktasını belirleyen bir yazılım geliştirilmiştir. Herhangi bir *i* anındaki çerçeve için kendisinden önceki ve sonraki beşer görüntü çerçevesi incelenmiştir. Eğer kendisinden önceki beş çerçeve pasif *(0)*, kendisinden sonraki beş çerçeve aktif *(1)* olarak işaretlenmiş ise *i* anı kelimenin başlangıç noktası olarak belirlenmiştir. İkinci durumda; eğer kendisinden önceki beş çerçeve aktif *(1)*,kendisinden sonraki beş çerçeve pasif *(0)* olarak işaretlenmiş ise *i* anı kelimenin bitiş noktası olarak belirlenmiştir. Kelimenin başlangıç ve bitiş noktalarını belirlemek için oluşturulan dizi ve bu dizi üzerinde gösterilen başlangıç ve bitiş noktaları Şekil 3.5'te gösterilmektedir.

Başlangıç-bitiş noktasına göre izole edilmiş kelimeye ait bütün açı değerleri için bu noktalar esas alınarak veri seti yeniden güncellenmiştir.

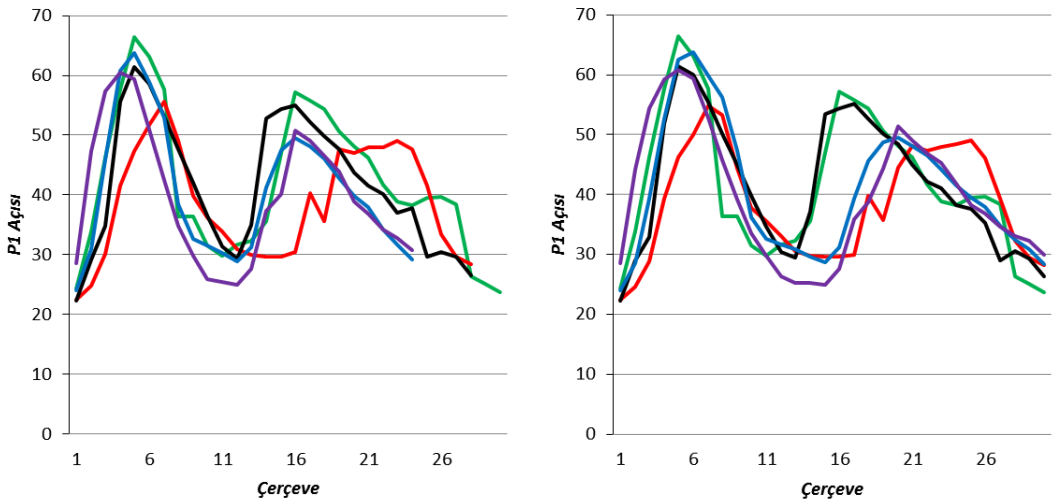


Şekil 3.5. Kelimenin başlangıç ve bitiş noktasının belirlenmesi

3.3.Kübik şerit interpolasyonu

Başlangıç ve bitiş noktalarına göre izole edilmiş bir kelimenin telaffuz uzunluğu genel olarak birbirinden farklılık göstermektedir. Yapılan deneyler sonucunda kelimelerin telaffuz sürelerinin eşit olmadığı belirlenmiştir. Bir kullanıcının aynı kelimeyi tekrar etmesi sırasında dahi kelimeyi telaffuz süresi birbirinden farklılık göstermektedir. Sınıflandırma işlemlerinde kullanılacak verinin eşit boyutlarda olması gerekmektedir. Her bir kelimeyi ifade eden çerçeve sayısı aynı uzunluğa getirilerek, kullanılacak matris temel bileşen analizi öncesi uygun hale getirilmiştir. Her kelimenin söyleniş uzunluğunu eşit hale getirmek için kübik şerit interpolasyonu *(Cubic Spline Interpolation)* yöntemi uygulanmıştır.

Şekil 3.6(a)'da bir kullanıcının "beyaz" kelimesini 5 farklı tekrarı süresince *P*1 açısında meydana gelen değişimler gösterilmektedir. Kullanıcının kelime telaffuzu sırasında başlangıç ve bitiş noktalarına göre işaretlenmiş bu verilerde görüldüğü gibi, kullanıcı bir kelimeyi 25 ile 30 çerçeve arasında söylemiştir. Kübik şerit interpolasyonu algoritması uygulandıktan sonra Şekil 3.6(b)'de gösterildiği gibi, veriler aynı çerçeve sayısına getirilmiştir.



Şekil 3.6. a) Kübik şerit interpolasyonu öncesi b) Kübik şerit interpolasyonu sonrası

Kübik şerit interpolasyonu çerçeveler arasındaki geçişlerde devamlılığı bozmadan veri setini tanımlanan aralığa getirmede kullanılan etkili bir yöntemdir. Bu işlem sonrasında her bir kelimeye ait vektör aynı uzunluğa getirilmiştir.

3.4.Verilerin belirlenen aralığa normalizasyonu

Yapılan çalışma neticesinde her kullanıcının konuşmaları sürecince aynı kelimeleri tekrarlarken bile $P1$ açısının sabit bir aralık içerisinde salınım yapmadığı görülmüştür.

Kişilerin fiziksel özellikleri ve konuşma alışkanlıkları nedeniyle kelime telaffuzları farklılık göstermektedir. Yapılan deneyler sonucunda, Şekil 3.7(a)'da bir kullanıcının "beyaz" kelimesini telaffuz ederken $P1$ açısı 22 derece ile 66 derece arasında salınım yaparken, Şekil 3.7(c)'de diğer kullanıcının aynı kelimeyi telaffuzu 33 derece ile 73 derece arasında olabilmektedir. Kullanıcıların aynı kelimeyi telaffuzlarında; kendi tekrarları ve diğer kullanıcıların telaffuzları arasında meydana gelen farklılıklar sınıflandırma başarısını düşürmektedir. Bu problemi ortadan kaldırmak için, kişilerin konuşma esnasındaki maksimum ve minimum dudak açıklıklarını normalize etmek gerekmektedir.

Normalizasyon işlemi, her kullanıcı ve kullanıcıya ait her açı için ayrı ayrı yapılmıştır. j'inci açının i'inci çerçevedeki değerini normalize etmek için, bütün açı değerlerinde kayıt süresince maksimum ve minimum değerler bulunup min_j ve max_j değişkenleri olarak kaydedilmiştir. Bulunan bu değerlere göre yapılan işlemde; Denklem (3.4) kullanılarak konuşma esnasındaki en büyük açı değeri 30 dereceye, en küçük açı değeri ise 0 dereceye karşılık gelecek şekilde açılar doğrusal olarak normalize edilmiştir.

$$g[k] = \frac{f\left(\frac{N}{30} * k\right) - min_j}{max_j - min_j} * 30 \tag{3.4}$$

$g[k]$ = Normalize edilmiş çerçevedeki açı sayısı

f = Kübik şerit interpolasyon fonksiyonu

N = Çerçeve sayısı

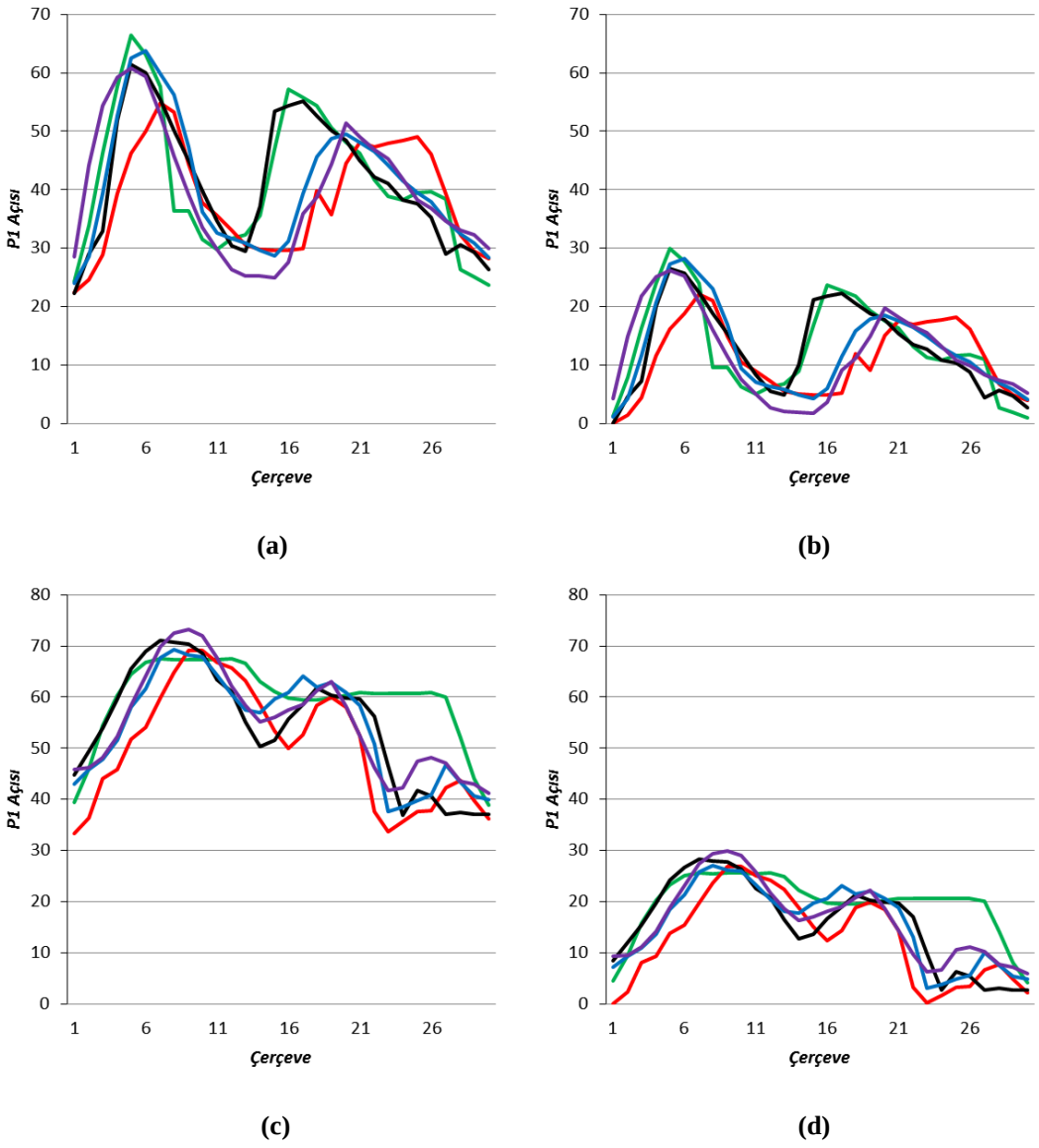
min_j= Kullanıcının j'inci açı için bütün kelimeleri telaffuzu sırasında açının sahip olduğu en küçük değeri

max_j= Kullanıcının j'inci açı için bütün kelimeleri telaffuzu sırasında açının sahip olduğu en büyük değeri

i = Çerçeve numarası $(i = 1,2, \dots, N)$

k = Normalize edilmiş çerçeve numarası $(x = 1,2, \dots, N)$

Çerçeve sayısı 30 olacak şekilde çerçeveler genişletilmiş veya daraltılmıştır. Bunun için f kübik şerit interpolasyonu kullanılmıştır.



Şekil 3.7. a) A kullanıcısının "beyaz" kelimesini 5 tekrarı b) A kullanıcısından elde edilen verilerin normalizasyon işlemi sonrası c) B kullanıcısının "beyaz" kelimesini 5 tekrarı d) B kullanıcısından elde edilen verilerin normalizasyon işlemi sonrası

3.5.Temel bileşen analizi

Büyük veri setleri genellikle birçok değişken üzerinde ölçü birimlerine sahiptir. Orijinal veri setinin sahip olduğu bilgileri koruyarak değişken sayısını düşürmek mümkündür. Boyut indirgeme problemi için birçok yöntem önerilse de temel bileşen analizi bu yöntemler arasında en yaygınıdır [40].

p rastgele değişkenlerinin x vektörü üzerinde n tane ölçümünde boyut sayısının p'den q'ya düşürülmesi hedeflenmektedir. Böylece eksenler birbirinden bağımsız hale getirilecektir. Bu nedenle, boyut indirgemek için temel bileşen analizi ile %95 oranında değişken varyansı kapsayacak şekilde veri seti 15 boyuta indirgenmiştir.

4. KELİME SINIFLANDIRMA SONUÇLARI ve TELAFFUZ KALİTESİ BELİRLEMEK İÇİN ÖZNİTELİKLERİN ÇIKARIMI

4.1.K-En Yakın Komşu Yöntemi ile Sınıflandırma

Bölüm 3'te tanımlanan yöntemler kullanılarak elde edilen veri setinin tutarlılığını test etmek için K-en yakın komşu algoritması ve Yapay Sinir Ağları kullanılmıştır. Ayrıca elde edilen açı değerleri için en iyi kombinasyonları elde etmek için veri setini oluşturan bütün açı değerlerinin alt kümeleri ile kombinasyonları oluşturulmuştur. Oluşturulan bu kombinasyonlar KNN algoritmasında Öklid uzaklığı kullanılarak test edilmiştir. 15 kelimenin birbirinden farklı 10 kullanıcı tarafından 5'er kez tekrarından yani 750 kelimeden meydana gelen veri seti kullanılarak sınıflandırma başarısı test edilmiştir.

En yakın komşu parametresi 1 olarak belirlenmiştir. Öklid ve Manhattan uzaklıkları başlangıç aşamasında test edilerek, Öklid uzaklığının daha iyi sınıflandırma başarısına sahip olduğu tespit edilmiş ve diğer bütün testlerde Öklid uzaklığı kullanılmıştır [41]. Sınıflandırmada eğitim için veri setinin %80'i kullanılmıştır, kalan %20'lik kısmı test için kullanılmıştır.

Öncelikle kişiye bağlı sınıflandırma başarımı test edilip, kullanıcıların konuşmalarını sınıflandırmada kullanılacak baskın açı kombinasyonları bulunmuştur. Bu test sonucunda her kullanıcının kendisine ait bir konuşma biçimi olması nedeniyle, baskın açıların farklılık gösterdiği sonucuna varılmıştır. Bu nedenle bütün kullanıcılar için ortak açı değerleri bulunmuştur.

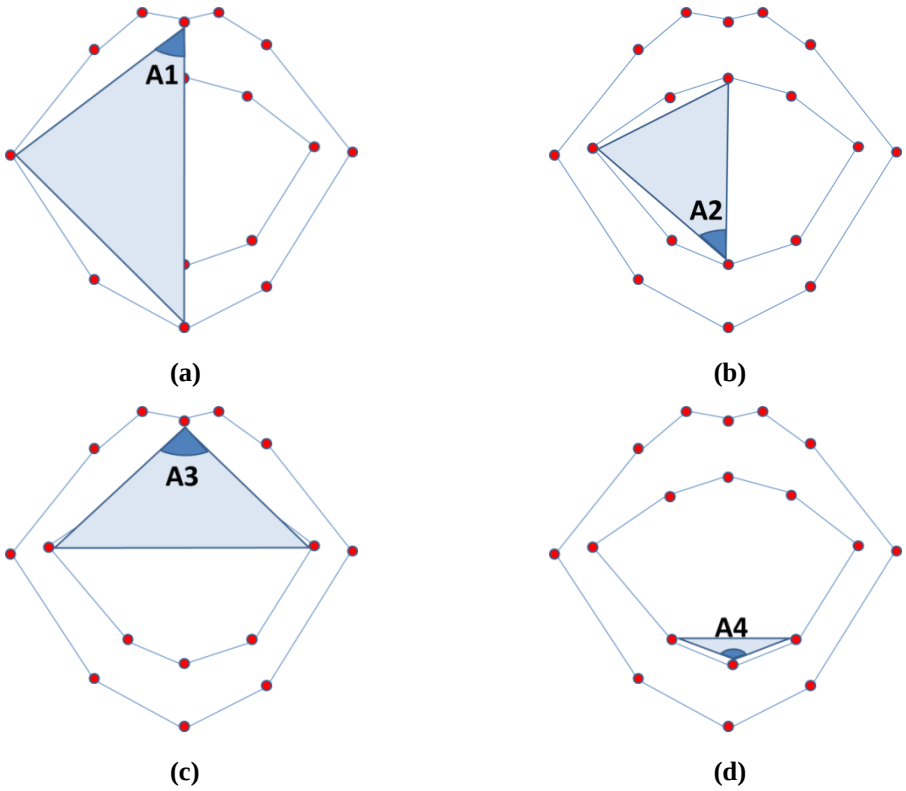
Test verisi seçilirken bütün kullanıcıların telaffuz ettiği bütün kelimelerden birer örnek alınmıştır. Sonuç olarak kullanıcının telaffuz ettiği bir kelimenin beş tekrarından dört tanesi eğitim, kalan bir tanesi ise test için kullanılmıştır.

Bölüm 3, Çizelge 3.1'de ifade edilen 21 açı değerinin kombinasyonundan meydana gelen bütün alt kümeleri test edildiğinde;

- 3 açı kullanılarak elde edilen en yüksek sınıflandırma başarısı %64,67'dir,
- 4 açı kullanılarak elde edilen en yüksek sınıflandırma başarısı %66'dir,

- 5 açı kullanılarak elde edilen en yüksek sınıflandırma başarısı %67,33'dür,
- 6 açı kullanılarak elde edilen en yüksek sınıflandırma başarısı %68'dir,
- 7 ve daha büyük elde edilen açı değerlerinin kombinasyonunda ise bu değer gittikçe düşmektedir.

Elde edilen sınıflandırma başarıları karşılaştırıldığında tam arama yöntemi ile en iyi sınıflandırma başarısı 6 açıdan meydana gelen kombinasyondan meydana gelmektedir. Fakat sonraki aşamada DTW yöntemi ile optimum eğrilme yolu hesaplanacağı için, açı sayısı arttığında işlem yükü de artmaktadır. Bu nedenle 4 açının meydana getirdiği kombinasyonu, benzerlik hesaplama işlemi için kullanılmıştır. Bu 4 açı Şekil 4.1'de, bu açılar meydana getiren vektörler ise Çizelge 4.1'de gösterilmektedir.



Şekil 4.1. En başarılı açı değerleri a) 17-12-9 noktalarından oluşan A1 açısı b) 1-7-3 noktalarından oluşan A2 açısı c) 1-12-5 noktalarından oluşan A3 açısı d) 18-17-16 noktalarından oluşan A4 açısı

Çizelge 4.1. En başarılı sınıflandırma sonucuna sahip 4 açıyı oluşturan vektörler

Açı No	Kullanılan Noktalar	v vektörü	w vektörü
A1	17-12-9	$v = (12,17)$	$w = (12,9)$
A2	1-7-3	$v = (7,1)$	$w = (7,3)$
A3	1-12-5	$v = (12,1)$	$w = (12,5)$
A4	18-17-16	$v = (17,18)$	$w = (17,16)$

Yapılan bu çalışma neticesinde bütün kullanıcılara ait en iyi sınıflandırma becerisin sahip öznitelik kombinasyonu belirlenmiştir. DTW algoritması iki zaman serisi arasındaki optimum uyumu bulur. Algoritmanın $O(N^2)$ zaman ve bellek karmaşıklığı nedeni ile en tutarlı sonucu veren 6 açıdan meydana gelen kombinasyon yerine 4 açı kullanılmıştır. Bu açılar kullanılarak %66 oranında sınıflandırma başarısı elde edilmiştir. Çizelge 4.2'de 4 açı kullanılarak, 15 kelimenin 10 farklı tekrarından meydana gelen test verisinin hata matrisi gösterilmektedir. Matrisin her bir sütunu tahmin edilen bir sınıftaki örnekleri temsil ederken, her bir satırı ise bir örneğin gerçek sınıfını ifade eder. Matristeki köşegende ise doğru sınıflandırılan örneklerin sayısı gösterilmektedir.

Elde edilen hata matrisi incelendiğinde beyaz, bordo, menekşe, siyah ve yeşil kelimeleri 10 örnekten 8 tanesini başarılı olarak sınıflandırmıştır. Mavi, lacivert ve pembe kelimelerinin sınıflandırma başarıları oldukça düşüktür. 10'ar kelimelik test örneklerinden yaklaşık olarak yarı yarıya doğru olarak sınıflandırıldığı görülmektedir. Bordo kelimesi ile turuncu kelimesi sınıflandırma işlemi sonrasında birbirleri yerine hatalı olarak sınıflandırıldıkları görülmüştür. Bunun nedeni iki kelimenin de söylenişlerinde dudağın aldığı şeklin birbirine benzerlik göstermesidir. Aynı şekilde turkuaz ve turuncu kelimeleri de söyleyiş benzerlikleri birbirine yakınlık gösterdiği için sınıflandırma işlemi sonucunda birbirleri yerine yanlış olarak sınıflandırılmışlardır. Söyleniş biçimleri birbirine benzerlik göstermeyen kelimelerin birbirleri yerine sınıflandırıldığı pembe ile sarının sınıflandırıldığı örneklerde ise, bu hatanın veri toplama işleminde meydana gelen tutarsızlıklar ve kelime telaffuzu sırasında meydana gelen öngörülmeyen hatalardan kaynaklandığı düşünülmektedir.

Çizelge 4.2. KNN ile 4 açı kullanılarak yapılan sınıflandırmanın hata matrisi

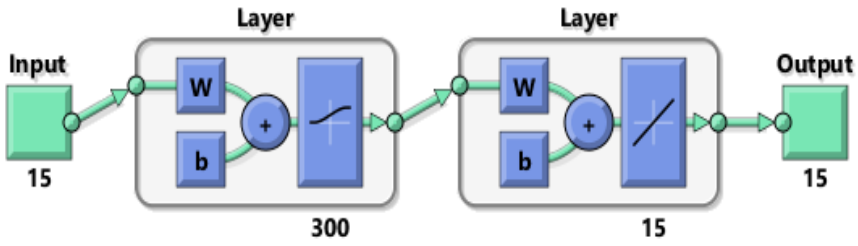
Hedeflenen Sınıf	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o				
a	8	1	0	0	0	0	0	0	0	0	0	0	0	0	1	|	a	=	Beyaz
b	0	8	0	0	0	0	0	0	0	0	0	0	0	2	0	|	b	=	Bordo
c	0	0	6	0	1	0	0	0	0	0	2	0	0	0	1	|	c	=	Gri
d	0	0	0	7	0	1	0	1	0	1	0	0	0	0	0	|	d	=	Kahverengi
e	0	2	0	0	6	0	0	0	0	0	0	0	0	1	1	|	e	=	Kırmızı
f	0	0	0	2	1	5	0	2	0	0	0	0	0	0	0	|	f	=	Lacivert
g	1	0	0	2	0	0	4	0	2	1	0	0	0	0	0	|	g	=	Mavi
h	0	0	0	0	0	0	0	8	0	0	2	0	0	0	0	|	h	=	Menekşe
i	0	1	0	0	0	0	0	0	6	2	0	0	0	1	0	|	i	=	Mor
j	0	0	0	0	0	0	1	1	0	5	2	1	0	0	0	|	j	=	Pembe
k	0	0	0	0	0	0	0	2	0	0	7	0	0	0	1	|	k	=	Sarı
l	0	0	1	0	0	0	0	0	0	0	1	8	0	0	0	|	l	=	Siyah
m	0	0	0	0	0	0	0	0	0	0	0	1	6	2	1	|	m	=	Turkuaz
n	0	1	0	0	0	0	0	0	0	0	0	0	1	7	1	|	n	=	Turuncu
o	0	0	1	0	0	0	1	0	0	0	0	0	0	0	8	|	o	=	Yeşil

Sonuç Sınıfı

4.2.Yapay Sinir Ağları İle Sınıflandırma

Elde edilen veri seti için Çizelge 3.1'de gösterilen 21 tane açının 30 fps ile alınan kayıtları kullanılmıştır. Veri setinin bir kelimeyi ifade eden tek satırı 21x30 yani 630 boyuttan meydana gelmektedir. Bir açı için kullanılan 30 çerçeveli görüntü, ANN sınıflandırma yönteminde ağ giriş verisi olarak yüksektir. Bu nedenle, boyut indirgemek için temel bileşen analizi ile %95 oranında değişken varyansı kapsayacak şekilde veri seti 15 özniteliğe indirgenmiştir.

Çok katmanlı ağların gizli katmanlarında tipik olarak sigmoid transfer fonksiyonu kullanılmaktadır. Bu fonksiyonlar genellikle sıkıştırma *(squashing)* fonksiyonlar olarak adlandırılmaktadır. Çünkü bu fonksiyonlar sonlu bir çıkış aralığı içine sonsuz bir giriş aralığı sıkıştırmaktadırlar. Girdi büyüdükçe, sigmoid fonksiyonların eğimi sıfıra yaklaşmaktadır. Bu durum, sigmoid fonksiyonlar ile çok katmanlı ağların eğitiminde dik iniş (*steepest descent)* kullanıldığında bir probleme neden olmaktadır; çünkü eğim *(gradient)* oldukça küçük değere sahip olmaktadır. Bu nedenle, ağırlıklar ve eşik değerleri optimal değerlerinden uzak olsa bile ağırlıklarda ve eşik değerlerinde küçük değişimlere neden olurlar.



Şekil 4.2. Yapay sinir ağı yapısı

Esnek geri yayılım algoritmasının *(Resilient backpropagation-Rprop)* amacı, eğimlerdeki kısmi türevlerin olumsuz etkilerini elimine etmektir. Ağırlıkların yönlerinin güncellenmesine karar vermek için sadece türevin işareti kullanılır, türevin eğiminin ağırlıkların güncellenmesi üzerinde herhangi bir etkisi yoktur. Ağırlık değişiminin büyüklüğü ayrı bir güncelleştirme değeri tarafından belirlenmektedir. Her ağırlık ve eşik değerinin güncellenme değeri ağırlık değişimi için artış *(delt_inc)* faktörü ile arttırıldığında, bu ağırlıklara bağlı olan performans fonksiyonlarının türevi iki başarılı iterasyon için aynı işarete sahip olmaktadır. Güncellenme değeri ağırlık değişimi için azalış *(delt_dec)* faktörü ile azaltıldığında, ağırlıklarla ilişkili olan türev önceki iterasyona göre işaretini değiştirmektedir. Eğer türev sıfıra eşitse, güncellenme değeri aynı kalır. Ağırlıklar salınım yaptığında, ağırlık değişimi azaltılır. Birden fazla iterasyon için ağırlık aynı yönde değişime devam ederse, ağırlığın eğimindeki değişim arttırılır [42].

Oluşturalan 2 katmanlı ileri beslemeli ağ yapısı Şekil 4.2'de gösterilmiştir. Transfer fonksiyonları olarak logaritmik sigmoid fonksiyonu ve doğrusal transfer fonksiyonları kullanılmıştır. Oluşturulan bu yapı için KNN yönteminde kullanılan eğitim ve test verilerinden yararlanılmıştır. Öğrenme kuralı olarak esnek geri yayılım *(resillient back propogation)* kullanılmıştır. Elde edilen sonuçlar eğitim başarısı %84, test başarısı %67.33 olarak belirlenmiştir.

Çizelge 4.3'te bileşen analizi ile %95 oranında değişken varyansı kapsayacak şekilde 15 özniteliğe indirgenmiş veri setinin, ANN ile sınıflandırılmasından elde edilen hata matrisi gösterilmektedir. Matrisin her bir sütunu tahmin edilen bir sınıftaki örnekleri temsil ederken, her bir satırı ise bir örneğin gerçek sınıfını ifade eder. Matristeki köşegende ise doğru sınıflandırılan örneklerin sayısı gösterilmektedir.

Çizelge 4.3. ANN ile yapılan sınıflandırmanın hata matrisi

Hedeflenen Sınıf	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	
a	8	0	0	0	0	0	0	1	0	1	0	0	0	0	0	a = Beyaz
b	0	7	0	0	0	0	0	0	0	0	0	0	1	2	0	b = Bordo
c	0	0	5	0	4	0	0	0	0	0	1	0	0	0	0	c = Gri
d	0	0	1	7	0	2	0	0	0	0	0	0	0	0	0	d = Kahverengi
e	0	1	0	2	6	0	0	0	0	0	0	0	1	0	0	e = Kırmızı
f	0	1	0	1	0	5	0	1	1	0	0	0	0	1	0	f = Lacivert
g	0	0	1	0	0	0	7	0	1	1	0	0	0	0	0	g = Mavi
h	0	0	1	0	0	0	0	8	0	1	0	0	0	0	0	h = Menekşe
i	0	1	0	0	0	1	0	0	7	0	1	0	0	0	0	i = Mor
j	0	0	0	1	0	0	0	0	0	5	3	0	0	1	0	j = Pembe
k	0	0	0	0	0	0	1	0	0	2	5	1	0	1	0	k = Sarı
l	0	0	0	0	0	0	0	1	0	0	1	7	0	0	1	l = Siyah
m	0	0	0	0	0	0	0	0	0	0	0	0	8	1	1	m = Turkuaz
n	0	1	0	0	1	0	0	0	0	0	0	0	1	7	0	n = Turuncu
o	0	0	1	0	0	0	0	0	0	0	0	0	0	0	9	o = Yeşil

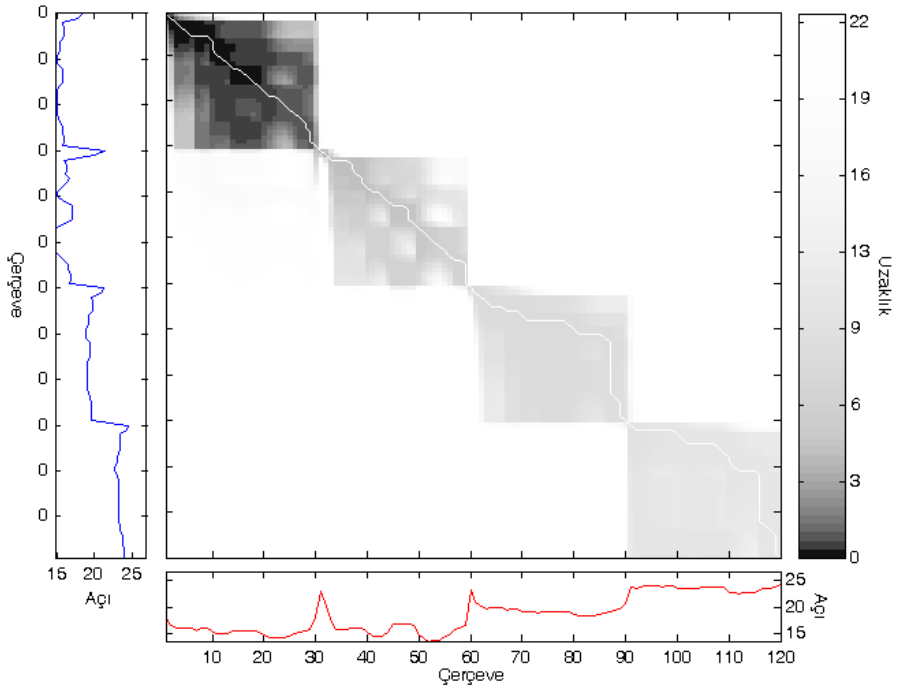
Sonuç Sınıfı

Elde edilen hata matrisi incelendiğinde beyaz, menekşe, turkuaz ve yeşil kelimeleri 10 örnekten 8 tanesini başarılı olarak sınıflandırmıştır. Lacivert, pembe ve sarı kelimelerinin sınıflandırma başarıları oldukça düşüktür. 10'ar kelimelik test örneklerinden yarı yarıya doğru olarak sınıflandırıldığı görülmektedir. Bordo kelimesi ile turuncu kelimesi sınıflandırma işlemi sonrasında birbirleri yerine hatalı olarak sınıflandırıldıkları görülmüştür. Bunun nedeni iki kelimenin de söylenişlerinde dudağın aldığı şeklin birbirine benzerlik göstermesidir. Aynı şekilde turkuaz ve turuncu kelimeleri de söyleyiş benzerlikleri birbirine yakınlık gösterdiği için sınıflandırma işlemi sonucunda birbirleri yerine yanlış olarak sınıflandırılmışlardır. Söyleniş biçimleri birbirine benzerlik göstermeyen kelimelerin birbirleri yerine sınıflandırıldığı pembe ile sarının birbiri yerine sınıflandırıldığı örneklerde ise, bu hatanın veri toplama işleminde meydana gelen tutarsızlıklar ve kelime telaffuzu sırasında meydana gelen öngörülmeyen hatalardan kaynaklandığı düşünülmektedir.

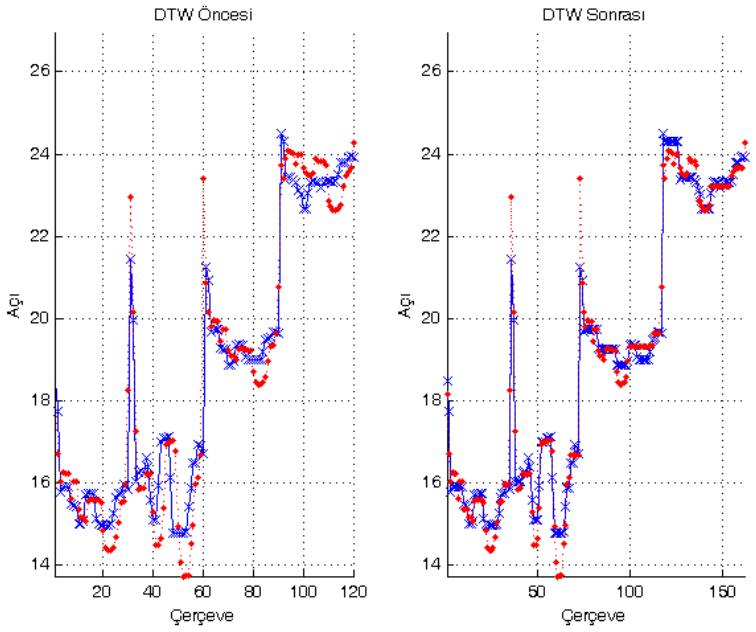
4.3.Telaffuz kalitesini belirlemek için iki telaffuzun benzerliğini hesaplamak

Engelli ya da konuşma terapisine ihtiyaç duyan bir kişinin dudak taklit yeteneğini geliştirmesi amacıyla, kamera karşısında farklı iki kişi tarafından telaffuz edilen bir kelimenin, kişiye ya da kelimeye bağımlılık olmadan, birbirine olan benzerliğini değerlendirecek bir sistem önerilmiştir. Bu sistemde ayırt edici öznitelik olarak KNN ile yapılan testler sonucunda elde edilen en iyi sınıflandırma becerisine sahip Çizelge 4.1'de gösterilen 4 açının oluşturduğu iki vektörün birbirine olan Öklid uzaklıkları kullanılmıştır.

İki kelimenin benzerliğini belirlemek için DTW algoritmasından yararlanılmıştır. Bölüm 2.5.3'te anlatılan DTW algoritması ile Şekil 4.3'te gösterilen, iki vektör arasındaki eğrilme matrisi ve optimum eğrilme yolu hesaplanmıştır. Optimum eğrilme yolu kullanılarak iki sinyal hizalanmıştır. Şekil 4.4(a)'da sinyalin başlangıç hali, Şekil 4.4(b)'de optimum eğrilme yoluna göre hizalanmış hali gösterilmektedir. DTW ile hizalanmış iki sinyal elde edildikten sonra, KNN yönteminde de kullanılan Öklid uzaklıkları hesaplanmıştır.



Şekil 4.3. Bir eğrilme matrisi ve optimum eğrilme yolu



(a) **(b)**

Şekil 4.4. Optimum eğrilme yolu kullanılarak sinyal hizalama işlemi a) İki sinyalin DTW önceki değerleri b) DTW işlemi uygulandıktan sonraki değerleri

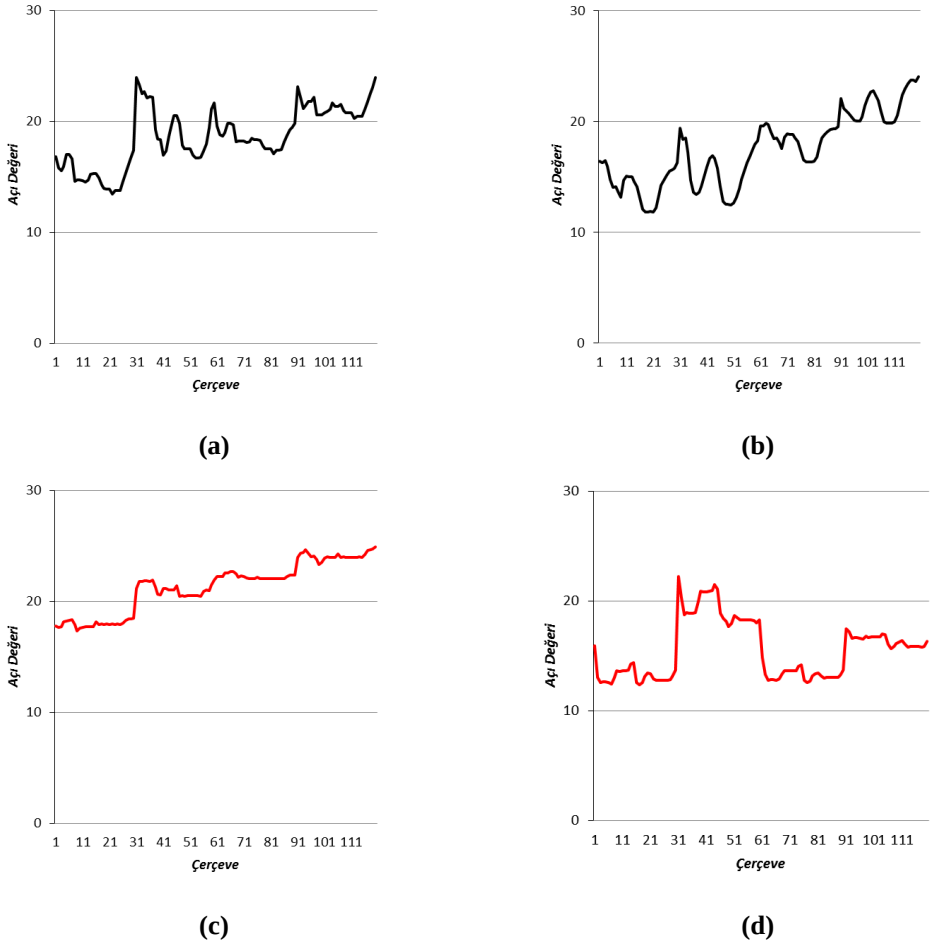
Kamera karşısındaki kişilerin aynı kelimeyi telaffuz etmeye çalıştıkları varsayılarak, iki kelime telaffuzunu ifade eden açı vektörlerinin Öklid uzakları oluşturulmuştur. Bu sistemde kullanılan vektörü elde etmek için Çizelge 4.1'de gösterilen KNN algoritması ile en iyi sınıflandırma başarısına sahip 4 açı birbirine eklenerek 120 çerçeveden oluşan tek bir vektör oluşturulmuştur. Bu vektörlerin birbirine olan Öklid uzakları deneysel olarak belirlenen eşik değerlerinden büyük ise telaffuzlar birbirine benzer olarak işaretlenmiştir.

Tanımlanan yönteme göre birbirine benzer olarak sınıflandırılan, "bordo" kelimesinin iki farklı tekrarı Şekil 4.5(a) ve (b)'de, benzer olarak sınıflandırılmayan iki farklı tekrar Şekil 4.5(c) ve (d)'de gösterilmiştir.

Çizelge 4.4. Bordo kelimesine ait 10 tekrarın karşılaştırılması ve birbirlerine olan benzerlikleri

Tekrar	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.
1.	1	1	1	1	1	0	0	1	1	0
2.	1	1	1	1	1	1	0	1	1	1
3.	1	1	1	0	0	1	0	1	1	0
4.	1	1	0	1	1	0	0	0	1	1
5.	1	1	0	1	1	0	1	0	1	1
6.	0	1	1	0	0	1	0	1	0	0
7.	0	0	0	0	1	0	1	0	1	1
8.	1	1	1	0	0	1	0	1	1	0
9.	1	1	1	1	1	0	1	1	1	1
10.	0	1	0	1	1	0	1	0	1	1

Çizelge 4.4'te 10 farklı kişi tarafından telaffuz edilen "bordo" kelimesinin birbirleri ile karşılaştırıldıklarında benzer (1) ya da değil (0) olarak üretilen sonuçlar gösterilmiştir. Karşılaştırma işlemi sonrasında 6. ve 7. tekrarların diğer tekrarlar ile karşılaştırıldığında, diğer 9 tekrarın 6'sında benzemez (0) olarak işaretlendiği görülmektedir. Bu tekrarlar KNN ile yapılan sınıflandırma sonucunda da başarısız olarak sınıflandırılmıştır. Fakat 1'inci tekrara bakıldığında da KNN ile sınıflandırılan bir kelimenin diğer 4 kelime ile benzemediği (0) görülmektedir. Veri setini oluşturan diğer kelimelerle de yapılan karşılaştırmalar sonucunda, kelime benzerlik testinin herhangi bir eğitim verisine sahip olmadan yaptığı karşılaştırmanın kabul edilebilir olduğu sonucuna varılmıştır.



Şekil 4.5. a) Benzer olarak sınıflandırılan 1. telaffuz b) Benzer olarak sınıflandırılan 2. Telaffuz c) Benzemez olarak sınıflandırılan 3. telaffuz d) Benzemez olarak sınıflandırılan 4. Telaffuz

5. SONUÇ VE ÖNERİLER

Bu çalışmada, konuşma terapisine ihtiyaç duyan kişilerin dudak taklit yeteneklerini kendi kendilerine analiz etmelerini sağlayan bir yöntem önerilmiştir. Bu yöntemi elde etmek için dudak hareket özelliklerini en iyi şekilde ifade eden özniteliklerin çıkarımı yapılmıştır.

Görüntü kayıt işlemi için MS Kinect kamerası kullanılmıştır. Bu kamera ile elde edilen veriler incelendiğinde kameranın dudağın anlık hareketlerindeki derinlik bilgisini yeterince hassas tespit edemediği görülmüştür. Ayrıca kayıt işlemi esnasında ortamdaki ışıklandırma ve kayıt ortamından güneş ışığını izole etmek verinin tutarlılığı açısından büyük önem taşımaktadır.

Sistemin başarısını test etmek ve telaffuz esnasında dudak hareketlerini en iyi şekilde ifade eden öznitelikleri belirlemek için 10 farklı kişinin 15 kelimeyi 5'er tekrarından meydana gelen 750 kelimelik bir veri seti oluşturulmuştur. MS Kinect kamerası tarafından elde edilen görüntüler analiz edilerek, görsel veri sayısal veriye çevrilip kayıt altına alınmıştır.

Oluşturulan test verisi kullanılarak KNN algoritması ile yapılan sınıflandırmada, 4 açı kullanılarak elde edilen sınıflandırma başarısı %66 oranındadır. Esnek geri yayılım *(resillient back propogation)* öğrenme kuralı kullanılarak oluşturulan ANN modeli ile yapılan sınıflandırmada eğitim başarısı %84, test başarısı %67.33 olarak belirlenmiştir.

Elde edilen sonuçlar doğrultusunda kelime telaffuzunu ifade eden en iyi açılar bulunmuştur. Bu açı değerlerinin oluşturduğu vektörler dinamik zaman bükmesi yöntemi ile hizalandıktan sonra benzerliklerini belirlemek için Öklid uzaklığından yararlanılmıştır.

Bir kelimenin iki farklı kullanıcı tarafından telaffuzunun benzerliklerini belirlemek için, görsel veriden elde edilen iki vektörün birbirine olan Öklid uzaklığı önemli ve ayırt edici bir özniteliktir. İki farklı kişinin aynı kelimeyi söylediği varsayımı ile yola çıkılarak oluşturulan telaffuz benzerliği bulma yönteminde, tanımlanan öznitelikler ile iki kelimenin birbirine benzer olup olmadığını belirlemeye yönelik önerilen çözüm, test verisi ile yapılan

sınıflandırmalarda doğru olarak sınıflandırılan örneklerin sonuçları ile benzerlik göstermiştir.

Dudak hareketleri konuşmanın algılanması ve yorumlanması esnasında önemli bir parametre olmasına rağmen sadece dudak hareketlerinden elde edilen öznitelikleri kullanarak tasarlanacak bir dudak okuma sisteminin başarım oranının ses ya da hem ses hem görüntü kullanılan sistemlere oranla daha düşük olacağı görülmüştür. Konuşma esnasında sesin oluşumu ses telleri ile başlar, ses yolunda, yutakta, ağızda ve burunda şekillenir. Diş, dil, sert ve yumuşak damak ile dudak kullanılarak ses şekil değiştirir. Dudağın oluşturduğu görsel veriler sesi oluşturan bütün bir sistemin parçasıdır ve ses ile birlikte kullanılarak birbirini tamamlayıcı olmalıdır.

KAYNAKLAR

[1] Sumby, W. H. ve Pollack, I., Visual contribution to speech intelligibility in noise, *The journal of the acoustical society of america*, 26**(212)** , 1954.

[2] Neely, K. K., Effect of visual factors on the intelligibility of speech, *The Journal of the Acoustical Society of America*, 28**(1275)** , 1956.

[3] Rabiner, L. ve Juang, B. H., Fundamentals of speech recognition, 1993.

[4] MacDonald, J. ve McGurk, H., Visual influences on speech perception processes. Perception & Psychophysics, 24(3), 253-257, 1978.

[5] Witt, S. M. ve Young, S. J., Phone-level pronunciation scoring and assessment for interactive language learning, Speech Communication, 30(2) , 95-108, 2000.

[6] Neumeyer, L., Franco, H., Digalakis, V. ve Weintraub, M., Automatic scoring of pronunciation quality, *Speech Communication*, 30**(2)** , 83-93, 2000.

[7] Cincarek, T., Gruhn, R., Hacker, C., Nöth, E. ve Nakamura, S., Automatic pronunciation scoring of words and sentences independent from the non-native's first language, *Computer Speech and Language***(23)** , 65-88, 2009.

[8] Turk, O. ve Arslan, L. M., "Pronunciation Scoring for the Hearing-Impaired," in SPECOM'2004:9. C*onference Speech and Computer*, 2004.

[9] Petajan, E., "Automatic Lipreading to Enhance Speech Recognition," in *Proceedings of Global Telecommunications Conference*, Atlanta, 265-272, 1984.

[10] Yau, W. C., Kumar, D. K., Arjunan, S. P. ve Kumar, S., "Visual speech recognition using image moments and multiresolution wavelet images," in *Computer Graphics, Imaging and Visualisation, 2006 International Conference on. IEEE*, 194-199, 2006.

[11] Liang, L., Liu, X., Zhao, Y., Pi, X. ve Vefian, A. V., "Speaker independent audio-visual continuous speech recognition," in *Multimedia and Expo, 2002. ICME'02. Proceedings. 2002 IEEE International Conference on. IEEE*, 25-28, 2002.

[12] Loy, G., Holden, E. J. ve Owens, R., "3D head tracker for an automatic lipreading system," in *Proc. Australian Conf. on Robotics and Automation (ACRA2000).*, 2000.

[13] Yoshinaga, T., Tamura, S., Iwano, K. ve Furui, S., "Audio-visual speech recognition using new lip features extracted from side-face images," in *COST278 and ISCA Tutorial and Research Workshop (ITRW) on Robustness Issues in Conversational Interaction.*, 2004.

[14] Iwano, K., Yoshinaga, T., Tamura, S. ve Furui, S., Audio-visual speech recognition using lip information extracted from side-face images, *EURASIP Journal on Audio, Speech, and Music Processing*, 1, 4-4, 2007.

[15] Myers, C. S. ve Habiner, L. F., A Comparative Study of Several Dynamic Time-Warping Algorithms for Connected-Word, *Bell System Technical Journal*, 60**(7)**, 1389-1409, 1981.

[16] Hinton, G. E. ve Salakhutdinov, R.R., "Reducing the dimensionality of data with neural networks," *Science*, 313, (5786), 504-507, 2006.

[17] Bagai, A., Gandhi, H., Goyal, R., Kohli, M. ve Prasad, T.V., "Lip reading using neural networks," *International Journal of Computer Science and Network Security*, 9, (4), 108-111, 2009.

[18] Yau, W. C., Kumar, D. K. ve Chinnadurai, T., "Lip reading technique using spatio-temporal templates and support vector machines," *Proceedings of the 13th Iberoamerican Congress on Pattern Recognition, Lecture Notes in Computer Science*, 5197, 610-617, 2008.

[19] Shin, J., Jin, L. ve Daijin, K., Real-time lip reading system for isolated Korean word recognition, *Pattern Recognition*, 44**(3)** , 559-571, 2011.

[20] Puviarasan, N. ve Palanivel, S., Lip reading of hearing impaired persons using HMM, *Expert Systems with Applications*, 38**(4)** , 4477-4481, 2011.

[21] Turk, O. ve Arslan, L. M., Speech recognition methods for speech therapy, In Signal Processing and Communications Applications Conference, 2004. Proceedings of the IEEE 12th, 410-413, 2004.

[22] Çetingül, H. E., Erzin, E., Yemez, Y. ve Tekalp, A., Multimodal

speaker/speech recognition using lip motion, lip texture and audio, *Signal processing*, 86**(12)** , 3549-3558, 2006.

[23] Kaynak, M. N., Zhi, Q., Cheok, A. D., Sengupta K., Jian, Z. ve Chung, K. C., Analysis of lip geometric features for audio-visual speech recognition, *IEEE Trans. Syst., Man, Cybern. A, Syst. Humans*, 34**(4)** , 564-570, 2004.

[24] Galatas, G., Potamianos, G. ve Makedon, F.,"Audio-visual speech recognition incorporating facial depth information captured by the Kinect," in *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European IEEE*, 2714-2717,2012.

[25] Yau, W. C., Kumar, D. K. ve Arjunan, S. P., Visual recognition of speech consonants using facial movement features., *Integrated Computer-Aided Engineering*, 14**(1)** , 2007, 49-61.

[26] Yau, W. C., Kumar, D. K., Arjunan, S. P. ve Kumar, S., "Visual speech recognition using image moments and multiresolution wavelet images.," in *IEEE*, 194-199, 2006.

[27] Zhang, J., Pan, P. ve Yan, Y., Automatic Scoring on English Passage Reading Quality, *Procedia Engineering* , 29, 2744-2748, 2012.

[28] Bernstein, J., Cohen, M., Murveit, H., Ritschev, D. ve Weintraub, M., "Automatic evaluation and training in English pronunication.," in *ICSLP'90*, 1185–1188, 1990.

[29] Eskenazi, M.," Detection of foreign speakers' pronunciation errors for second language training– preliminary results," in *ICSLP'96*, 1996.

[30] Arias, J. P., Yoma, N. B. ve Vivanco, H., Automatic intonation assessment for computer aided language learning, *Speech Communication***(52)** , 254-267, 2010.

[31] Kumar, K., Tsuhan, C. ve Stern, R. M., "Profile view lip reading," Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on. IEEE, (4), 429-432, 2007.

[32] Catuhe, D., Programming with the Kinect for Windows Software Development Kit.: O'Reilly Media, Inc., 2012.

[33] Elmas, Ç., *Yapay Sinir Ağları*: Seçkin Yayıncılık, 2003.

[34] Öztemel, E., *Yapay Sinir Ağları*: Papatya Yayıncılık, 2003.

[35] Anderson, D. ve George, M., *Artificial neural networks technology*: Kaman Sciences Corporation, 1992.

[36] Meinard, M., "Dynamic Time Warping," in *Information retrieval for music and motion*: Springer, 69-74, 2007

[37] Theodoridis, S. ve Koutroumbas, K., *Pattern Recognition*,Third Edition: Elsevier, 2006.

[38] Salvador, S. ve Philip, C., "Toward accurate dynamic time warping in linear time and space," *Intelligent Data Analysis*, 11, (5), 561-580, 2007.

[39] Bemdt, D. J. ve Clifford, J., Using Dynamic Time Warping to Find Patterns in Time Series, *KDD workshop*, 10**(16)** , 1994.

[40] Jolliffe, I., "Principal component analysis," in *Encyclopedia of Statistics in Behavioral Science*, 3rd ed.: John Wiley & Sons, Ltd, 2005, pp. 1580-1584.

[41] Yargıç, A. ve Doğan, M., "A lip reading application on MS Kinect camera.," in In Innovations in Intelligent Systems and Applications (INISTA), 2013 IEEE International Symposium IEEE., 2013, pp. 1-5.

[42] Demuth, H. ve Beale,M., Neural Network Toolbox For Use with MATLAB, 4th ed., 2002.