Düzce Üniversitesi Bilim ve Teknoloji Dergisi

Araştırma Makalesi

Sessizliğin Kaldırılması ve Konuşmanın Parçalara Ayrılması İşleminin Türkçe Otomatik Konuşma Tanıma Üzerindeki Etkisi

Saadin OYUCU [a,*], Hüseyin POLAT [a], Hayri SEVER [b]

[a] *Bilgisayar Mühendisliği Bölümü, Teknoloji Fakültesi, Gazi Üniversitesi, Ankara, TÜRKİYE*
[b] *Bilgisayar Mühendisliği Bölümü, Mühendislik Fakültesi, Çankaya Üniversitesi, Ankara, TÜRKİYE*
** Sorumlu yazarın e-posta adresi: saadinoyucu@gazi.edu.tr*
DOI: 10.29130/dubited.560135

ÖZET

Otomatik Konuşma Tanıma sistemleri temel olarak akustik bilgiden faydalanılarak geliştirilmektedir. Akustik bilgiden fonem bilgisinin elde edilmesi için eşleştirilmiş konuşma ve metin verileri kullanılmaktadır. Bu veriler ile eğitilen akustik modeller gerçek hayattaki bütün akustik bilgiyi modelleyememektedir. Bu nedenle belirli ön işlemlerin yapılması ve otomatik konuşma tanıma sistemlerinin başarımını düşürecek akustik bilgilerin ortadan kaldırılması gerekmektedir. Bu çalışmada konuşma içerisinde geçen sessizliklerin kaldırılması için bir yöntem önerilmiştir. Önerilen yöntemin amacı sessizlik bilgisinin ortadan kaldırılması ve akustik bilgide uzun bağımlılıklar sağlayan konuşmaların parçalara ayrılmasıdır. Geliştirilen yöntemin sonunda elde edilen sessizlik içermeyen ve parçalara ayrılan konuşma bilgisi bir Türkçe Otomatik Konuşma Tanıma sistemine girdi olarak verilmiştir. Otomatik Konuşma Tanıma sisteminin çıkışında sisteme giriş olarak verilen konuşma parçalarına karşılık gelen metinler birleştirilerek sunulmuştur. Gerçekleştirilen deneylerde sessizliğin kaldırılması ve konuşmanın parçalara ayrılması işleminin Otomatik Konuşma Tanıma sistemlerinin başarımını artırdığı görülmüştür.

Anahtar Kelimeler: *Otomatik konuşma tanıma, Sessizliğin kaldırılması, Konuşmanın parçalanması*

The Effect of Removal the Silence and Speech Parsing Processes on Turkish Automatic Speech Recognition

ABSTRACT

Automatic Speech Recognition systems are mainly developed using acoustic information. Paired speech and text data are used to obtain phoneme information from acoustic information. The acoustic models trained with these data cannot model all acoustic information in real life. For this reason, it is necessary to carry out certain pre-processing and eliminate the acoustic information that will reduce the performance of automatic speech recognition systems. In this study, a method for removing silences in the speech was proposed. The aim of the proposed method is to eliminate silence and to break down conversations that give long dependencies. The speech information, which does not contain any silence and is divided into pieces, is given as an input to the Turkish Automatic Speech Recognition system. In the output of the Automatic Speech Recognition system, the speech that is given as input to the system are presented by combining the corresponding texts. In the experiments carried out, it was seen that the removal of silence and parsing of speech increased the performance of Automatic Speech Recognition systems.

Keywords: *Automatic speech recognition, Silence removal, Speech parsing*

Geliş: 03/05/2019, Düzeltme: 09/05/2019, Kabul: 05/08/2019

I. GİRİŞ

Otomatik Konuşma Tanıma (ASR: Automatic Speech Recognition), konuşma fonksiyonlarından gelen fonem (phoneme) veya fonem dizilerini otomatik olarak yazıya çevirebilen akıllı bir sistem olarak tanımlanmaktadır [1]. Diğer bir ifade ile ASR, insanlar tarafından konuşulan kelimeleri mikrofon veya telefon girişi yoluyla bilgisayar tarafından okunabilir metne dönüştürmesini sağlayan bir teknoloji olarak ifade edilmektedir [2]. ASR sistemleri; telefon rehberi, veritabanı sorgulama uygulamaları, kimlik tanıma uygulamaları, tıp alanında konuşmaya yardımcı uygulamalar ve yabancı dile çeviri gibi çeşitli uygulama alanlarında yenilikler sunmaktadır [3]. ASR sistemleri akıllı ev sistemleri, sesli komutlar ile sağlanan güvenlik sistemleri, eğitim sistemleri ve Sesli Yanıt Sistemi (IVR: Interactive Voice Response) gibi birçok alanda yaygın olarak kullanılmaktadır.

ASR sistemleri farklı disiplinlerin bir arada kullanılması ile geliştirilmektedir. Ayrıca ASR'nin genel yapısı da oldukça karmaşıktır [4]. Bu nedenle ASR sistemleri üzerine birçok çalışma yapılsa da henüz istenilen başarım seviyesine ulaşılamamıştır. Teknolojinin ilerlemesi ile desen eşleştirme, işaret işleme, fonetik, doğal dil işleme ve bilgi teorisi alanlarında elde edilen kazanımlar ASR sistemlerinin başarımının artmasında önemli bir etken olmuştur. Son yıllarda ise özellikle hızlı, yüksek kapasiteli işlemci ve bellek teknolojilerinin yaygınlaşması, performanslarının sürekli artması ASR sistemleri için gerekli olan karmaşık ve güçlü modellerin oluşturulmasına olanak sağlamıştır [5].

ASR sistemlerinde temel olarak iki farklı model kullanılmaktadır. Bunlardan ilki akustik bilginin elde edildiği akustik modeldir [6]. İkincisi ise akustik bilgiyi desteklemek ve kullanılan dil yapısını matematiksel olarak ifade etmek için kullanılan dil modelidir [7]. Akustik modelde temel olarak, belirli zaman sinyali çerçevesindeki foneme ait sonsal olasılık hesaplanmaktadır. Fonem sırası belirlemek için genellikle Saklı Markov Modelleri (HMM: Hidden Markov Model) kullanılmaktadır [8]. Fonemlerin hizalanması Gauss Karışım Model (GMM: Gaussian Mixed Model) dağılımını kullanarak (HMM/GMM) ile elde edilmektedir [9]. HMM'de, durumlar arasındaki geçiş sırasında Markov varsayımları yapılıp bir durumdan başka duruma geçişte sadece bir önceki durum göz önünde bulundurulmaktadır. Bu kısıtlama uzun bağımlılığı olan dizilerin modellenmesini zorlaştırmaktadır. Ayrıca, HMM'deki durumların tahmin olasılığı da birbirinden bağımsızdır.

Birbirinden bağımsız tahmin olasılıklarını modellemenin bir diğer yaklaşımı da yapay sinir ağlarıdır. Yapay sinir ağındaki amaç bilinmeyen bir fonksiyonun çıkışlarını üretebilmektir. ASR'de bu çıkış fonem dizisini temsil etmektedir. Çok katmanlı yapay sinir ağı tabanlı akustik modelde ise fonemlerin sonsal olasılığı her bir pencere için bağımsızdır [10]. Bu bağımsızlık kelimede bulunan fonemlerin birbirinden bağımsız olması anlamına gelmektedir. Ancak bu varsayım fonemler arasında istatistiksel bağımlılıkları göz ardı etmektedir. Bu nedenle akustik modele girdi olarak verilecek bir konuşmanın gereksiz bilgi içermemesi gerekmektedir. Fonemlerin birbirinden bağımsız olması ve uzun bağımlılığın olmaması gerekir.

Literatür incelendiğinde gürültü seviyesi tahmin için dört farklı yöntemin kullanıldığı ve bu yöntemlerin temel olarak gürültü seviyesi tahmini yaparak sessizlik bilgisini elde etmek için kullanıldığı belirtilmiştir. Bu yöntemler gürültü spektrumları, Hirsch histogramları, ağırlıklı ortalama yöntemi ve düşük enerjili paket takibidir [11]. Literatürde gürültü enerjisinin hesaplanması ile konuşulan bölümler sırasındaki gürültü seviyesi tahmini yapılabildiği gösterilmiştir. Chen ve arkadaşları sessizlik bilgisini olasılık temelli ele alarak farklı bir bakış açısı sunmuştur [12]. Konuşma tanımadaki telaffuz olasılıkları ve kelimeye özgü sessizlik olasılıklarını modellenmişlerdir. Her bir kelimeyi takip eden sessizlik olasılığını modellemenin yanı sıra, her bir kelimenin sessizlikten sonra ortaya çıkma olasılığını da modellemişlerdir. Olasılık temelli yaklaşımın büyük veri setlerinde başarım sağladığını açıkça göstermişlerdir.

Konuşma tanıma sistemlerinde gürültünün ve sessizliğin sistem performansı üzerindeki etkisini azaltmak için çok sayıda gürültü azaltma ve konuşma etkililiğini tanıma tekniği geliştirilmiştir. Bu tekniklerde genellikle bir ses aktivitesi detektörü ile elde edilen gürültü istatistiklerinin tahmini

kullanılmıştır. Konuşma veya sessizlik tespiti, konuşma işlemede üzerine çalışılması gereken bir problemdir ve sağlam konuşma tanıma [13-15] sürekli olmayan konuşmanın iletimi [16], gerçek zamanlı algılama gibi birçok uygulamayı doğrudan etkilemektedir. Telefon konuşmalarında gürültü azaltma ve yankı iptal uygulamaları [17-18] bu alanda yapılan önemli çalışmalardandır. Konuşma ve konuşma dışı olguların sınıflandırma görevi için hazırlanan algoritmaların çoğu arka plan gürültüsü seviyesi arttığında başarısız olmaktadır. Son yıllarda çok sayıda araştırmacı, gürültülü bir sinyale ilişkin konuşmayı tespit etmek için farklı stratejiler geliştirmiştir [19-21]. Konuşma bilgisinin elde edilmesi yaklaşımlarının çoğu, gürültüye dayanıklı özelliklerin ve karar kurallarının türetilmesi için sağlam algoritmaların geliştirilmesi üzerine odaklanmıştır [22-23]. Farklı konuşma bilgisi tespit etme yöntemleri arasında enerji eşikleri [22], adım tespiti [24], spektrum analizi [25] sıfır geçiş oranı, periyodiklik ölçüsü esas alınmaktadır [16]. Literatür taramasında konuşma aktivitesinin tespitindeki ana zorluklara değinilmiştir. Ancak farklı çalışmalarda konuşma aktivitesinin başlangıç ve bitiş noktalarının tespiti ele alınmıştır. Li ve arkadaşları son nokta tespiti için spektrum analizi yerine özel bir filtre artı üç durumlu karar ağacı kullanma konusunda yeni bir yaklaşım önermişlerdir [26]. Filtre, tespitin doğruluğunu ve sağlamlığını sağlamak için çeşitli kriterler altında tasarlanmıştır. Tespit edilen uç durum daha sonra eş zamanlı olarak enerji normalleşmesine uygulanmıştır. Değerlendirme sonuçları, önerilen algoritmanın test edilen veri setinde hata oranlarının önemli ölçüde azalttığını göstermektedir.

Literatürde de görüldüğü gibi konuşma etkinliği, sessizlik bilgisi veya gürültü tanıma için birçok yaklaşım ve değerlendirme çerçevesi sunulmuştur. Fakat bu alan ile ilgili farklı çözümler problemi kapsamlı bir yaklaşım ile ele alsa da istenilen başarım sağlanamamıştır. Daha sağlam ve başarımı yüksek ASR sistemleri için konuşma etkinliği tanıma uygulamalarının başarımının arttırılması gerekir. Bu nedenle çalışma kapsamında sessizlik içeren bir konuşma içerisinde fonemlerin birbiri arasında bağımsız olmasının sağlanması, sadece konuşma içeren kısımların akustik modele girdi olarak verilmesi üzerine bir yöntem sunulmuştur. Ayrıca konuşmayı parçalara ayırarak uzun bağımlılıkların önüne geçilmesi amaçlanmıştır. Konuşma sinyalinin ele alındığı ilk andan itibaren öncelikle konuşmanın başlaması beklenmiştir. Geliştirilen sistem konuşma başladığı anda aktif hale gelmektedir. Sistemin aktif hale geldiği an parçalama işleminin başladığı noktadır. Konuşma bir sessizlik ortamı ile karşılaşıncaya kadar konuşma parçasının sürekliliği devam eder. Sessizlik bilgisi ile karşılaşıldığı an ise konuşma parçasın bitiş noktasıdır. Geliştirilen sistemde sessizlik algılama işleminin başlangıç noktası, sinyal büyüklüğünün sessizlik eşik değerinin altına düştüğü noktadır. Bitiş noktası ise sinyal büyüklüğünün sessizlik eşik değerin üstüne çıktığı nokta olarak ifade edilmiştir. Sessizlik ile karşılaşıldığı ilk anda konuşma hemen parçalanmamıştır. Bunun nedeni ise bazı konuşmacıların sakin, aralıklı ve durağan konuşmasındaki doğal sessizliklerin gerçekte konuşmanın özelliğinde var olmasıdır. Bu nedenle sessizlik ile karşılaşıldığı ilk andan itibaren bir zaman sayacı tutulmuştur. Zaman sayacının belirli bir değerin altında kalması durumunda konuşmanın parçalanması gerçekleştirilmemiştir. Böylelikle konuşmanın sürekliliği koruma altına alınmıştır.

Önerilen yöntem kullanılarak elde edilen konuşma parçaları ayrı birer girdi olarak sıralı şekilde ASR sistemine verilmiştir. Ayrı ayrı girdi olarak verilen konuşma parçalarına karşılık gelen metin verisi birleştirilerek sunulmuştur. Yapılan deneylerde bu yaklaşımın ASR sisteminin başarımını arttırdığı açıkça görülmüştür. Ancak bazı durumlarda ASR başarımına olumsuz etki ettiği de belirlenmiştir. Başarıma olumsuz etki etmesinin en büyük nedeni ise çok kısa konuşma parçalarından istenilen konuşma özellik bilgilerinin elde edilememesidir.

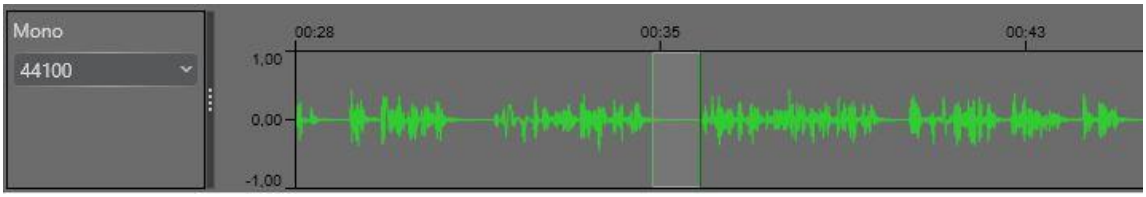
II. MATERYAL VE METOD

Dil bilimsel olarak anlamda değişme olmadan birbirleri yerine geçebilen bir dizi ses birimini ifade eden fonemlerden oluşan bir konuşmanın kayıpsız şekilde parçalara bölünmesi gerekir [27]. Kayıp ifadesi konuşmayı tamamlayıcı her türlü olgu fonem, harf, hece veya kelime bütünlüğünün kaybolması olarak ifade edilebilir. Herhangi bir fonem, harf veya hece kaybı ASR sisteminin çıkışında elde edilecek bilginin doğruluğunu etkileyecektir. Bu nedenle konuşma bilgisi doğru zamanda parçalara ayrılmalıdır. Kullanılacak materyal ve metodlar bu açıdan önemlidir ve dikkatli bir şekilde ele alınmalıdır. Bu

bölümde konuşma içerisindeki sessizliğin algılanması, konuşmanın parçalara ayrılması ve ASR sisteminin geliştirilmesinde kullanılan veri seti de dâhil olmak üzere bütün materyal ve metodlar detaylı olarak açıklanmıştır.

A. SESSİZLİĞİN ALGILANMASI VE KONUŞMA PARÇALARININ ELDE EDİLMESİ

Sessizlik bir ortamdaki algılanabilir sesin yokluğu anlamına gelmektedir. Fakat bazı durumlarda sessizlik kavramı değişerek sadece duyulabilir değil aynı zamanda insanlar tarafından duyulamayan sesler olarak ifade edilmektedir [28]. Literatürde gürültü seviyesi tespiti ile sessizlik bilgisinin algılanabileceği açıklanmaktadır. Bu işlem için genel olarak dört farklı yöntem kullanılmaktadır. Bu yöntemlerden gürültü spektrumlarının hesaplanması uzun konuşma ifadelerinde çok fazla hesaplama gücü gerektirmektedir. Hirsch histogramları, ağırlıklı ortalama yöntemi ve düşük enerjili paket takibi yöntemleri ise yüksek örneklem frekansına sahip sistemler yerine daha düşük örneklem frekansına sahip (örneğin telefon konuşma kayıtları v.b.) verilerde başarım göstermektedir. Bu nedenle çalışma kapsamında konuşma aktivitesinin tespiti farklı bir yaklaşım ile ele alınmış ve ASR sistemi üzerine etkisi araştırılmıştır. Belirtilen yaklaşımın ana çıkış noktası her ses veya konuşma sinyalinin belirli bir frekans bandında ilerlemesidir. Konuşma bilgisi belirli bir şiddete, frekansa ve genliğe sahiptir. Şekil 1'de bir konuşma dosyasının zaman–genlik dalga formu gösterilmiştir.



Şekil 1. *Konuşma dosyasının zaman-genlik dalga formu gösterimi*

Şekil 1'den yararlanılarak bir ses dosyası içerisindeki sessizliklerin bulunabilir olduğu anlaşılmaktadır. Dalga formu üzerinde yer alan örneklem frekansı göz önüne alındığında belirli bir frekansın altı sessizlik olarak nitelendirilebilir. Bu nedenle önerdiğimiz yöntemde ses dosyasının örneklem frekansına göre bir hesaplama yapılmıştır (Denklem 1).

$$Eşik\ Değeri = \frac{Örneklem\ Frekansı}{1000} \quad (1)$$

Denklem 1'de görüldüğü gibi sessizlik ifadesinin bulunması amacıyla örneklem frekansının belirli bir oranı alınmaktadır. Böylelikle elde edilen frekans bilgisinin aşağısında kalan değerler sessizlik olarak algılanmıştır. Denklem 1'de belirtilen oran değiştirilerek farklı amaçlar için kullanılabilir. Gerçekleştirdiğimiz deneylerde bu oranın en uygun sonucu verdiği görülmüştür.

Bu yöntemde ele alınması gereken temel sorun durağan ve aralıklı konuşmalarda sessizlik bilgisinin elde edilmesidir. Bu sorunu çözebilmek için sessizlik bilgisi ile karşılaşıldığı ilk andan itibaren arka planda bir zaman sayacı tutulmaktadır. Bu sayaç sessizlik bilgisi boyunca değiştirilmektedir. Sayaç belirli bir eşik değerine ulaştığı anda sessizlik bilgisinin var olduğu sonucuna varılmaktadır. Böylelikle durağan konuşan, hece veya harf kayıplarının yaşandığı konuşmaları içeren konuşmalarda sessizlik bilgisi başarılı bir şekilde elde edilmiş olur.

Belirtilen yöntemdeki ana amaç başlangıçtan itibaren sessizlik bilgisi ile karşılaşıldığı ana kadar ki konuşmanın ana konuşma içerisinden ayrılmasıdır. Ana konuşma dosyasından ayrılan her bir parçaya benzersiz bir kimlik verilerek geçici bellekte tutulması sağlanmıştır. Kimlik bilgisi aynı zamanda hangi ses parçasının hangi sırada olduğu bilgisini de tutmaktadır. Böylelikle parçalar arasında bir bütünlük sağlanmıştır.

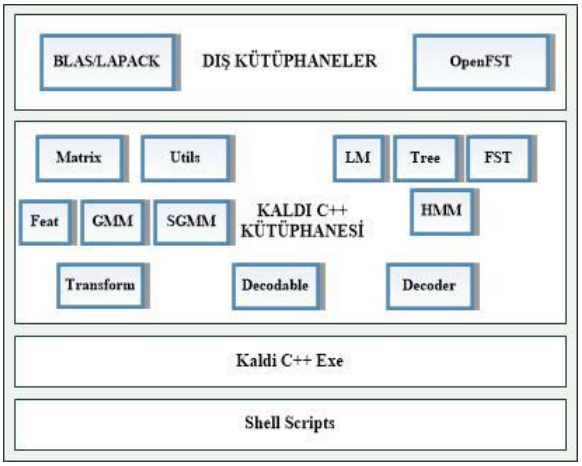
B. TÜRKÇE OTOMATİK KONUŞMA TANIMA SİSTEMİ

ASR sistemlerinde temel olarak iki farkı model kullanılmaktadır. Akustik ve dil modellemenin görevleri birbirinden farklıdır. ASR sistemlerinde kullanılan akustik modelleme ve dil modelleme işleminin bölünmesi, istatistiksel konuşma tanıma sistemlerinde denklem 2'deki gibi tanımlamaktadır [29].

$$W = \arg max P(W \mid A) = \arg max \frac{P(W)P(A|W)}{P(A)} \quad (2)$$

Denklem 2'de, akustik fonem veya özellik vektör dizisi X = X1, X2, … , Xn için konuşma tanıma sisteminde karşılık gelen kelime dizisini vermektedir. Denklem 1'de ifade edilen maksimum arka (posterior) olasılık P(W|X) değerine sahiptir. Denklemin en üst düzeye çıkarılması sabit X gözlemiyle gerçekleştirildiğinden payın maksimize edilmesine doğrudan bağlıdır. P(W) ve P(X|W) bileşenleri dil modellemesi ve akustik modelleme tarafından hesaplanan olasılıksal büyüklükleri oluşturmaktadır. Bu olasılık değerlerine göre çıkış parametreleri hesaplanmaktadır.

Bu çalışma kapsamında sessizlik bilgisinin algılanması ile parçalara ayrılan ses dosyalarının metne aktarılması için Türkçe ASR sistemine verilmesi gerekir. Bu nedenle temel bir Türkçe ASR sistemi geliştirilmiştir. ASR sistemi geliştirilirken Kaldi araç setinden faydalanılmıştır. Kaldi, C ++ ile yazılmış ve "Apache License v2.0" altında lisanslanan konuşma tanıma uygulamaları için kullanılan açık kaynaklı araç setidir [30]. Şekil 2'de Kaldi araç setine şematik bir bakış verilmiştir.



Şekil 2. *Kaldi araç seti [30]*

Kaldi araç seti, temel olarak iki harici kütüphaneye bağlıdır. Bunlardan ilki sonlu durum çerçevesi için kullanılan "OpenFst", diğeri ise sayısal cebir kütüphanesidir. Sayısal cebir kütüphanesi "BLAS" ve "LAPACK" olarak ikiye ayrılmıştır. Kaldi kütüphane modülleri, her biri harici kütüphanelerden sadece birine bağlı olacak şekilde konumlandırılmıştır. Kütüphane işlevlerine erişim C++ dilinde yazılmış kod parçacıkları ile sağlanmaktadır. Kaldi'de hazırlanan kod parçacıkları ve kütüphaneler konuşma tanıma sistemini oluşturmak ve çalıştırmak için betik dili tarafından çağrılmaktadır.

B. 1. Özellik Çıkarımı

Özellik çıkarımı, konuşma tanımanın en önemli adımlarındandır. Özellik çıkarımı bir konuşmayı farklı diğer konuşmalardan ayırmak için önemli rol oynamaktadır. Konuşma eyleminin doğası gereği her konuşmanın, konuşma bilgisi içerisine entegre edilmiş farklı bireysel özellikler bulunmaktadır [31]. Bu özellikler, ASR sistemleri için önerilen ve başarılı bir şekilde kullanılan farklı özellik çıkarım teknikleriyle elde edilmektedir. Özellik çıkarımı ve dalga formunu okuyabilmek için Kaldi, standart Mel Frekanslı Cepstral Katsayıları (MFCC: Mel-Frequency Cepstrum Coefficient) özelliklerinin oluşturulmasını desteklemektedir [32]. MFCC hesaplamanın tekniği temel olarak kısa vadeli analize dayanmaktadır. Bu nedenle her çerçeveden bir MFCC vektörü hesaplanmaktadır.

Cepstral katsayıları çıkarmak amacıyla konuşma örneği giriş olarak alınmakta ve sinyalin süreksizliğini en aza indirmek için Hamming penceresi uygulanmaktadır. Bu pencereler daha sonra Mel filtre bankasını oluşturmak için Ayrık Fourier Dönüşümü (DFT: Discrete Fourier Transform) ile birlikte kullanılmaktadır. Mel frekans eğrisine göre filtrelerin genişliği değişmekte ve böylece merkez frekansı etrafındaki kritik bant üzerindeki özellikler hesaplanmaktadır [33]. Kaldi, en çok kullanılan özellik çıkarım teknikleri için standart özellikleri (Cepstral sayıları, en düşük ve en yüksek frekans kesmeleri vb.) ön ayarlı bir şekilde sunmaktadır [34].

B. 2. Akustik ve Dil Modelleme

Bir konuşma işleminin akustik modellenmesi, tipik olarak konuşma dalga formundan hesaplanan özellik vektör dizileri için istatistiksel bilgilerin oluşturulması işlemidir [35]. Kaldi, akustik modelleme için birçok alt yapıyı desteklemektedir. Klasik Gauss Karışım Model (GMM: Gaussian Mixed Model) ve Alt Uzay Gauss Karışım Modelleri (SGMM: Subspace Gaussian Mixed Model) Kaldi ile rahatlıkla geliştirilebilmektedir [7]. Bireysel Gauss yoğunluklarını ayrı ayrı göstermek yerine, doğrudan doğal parametrelerle GMM sınıfını uygulanmaktadır.

GMM tabanlı akustik modelde fonem bağlamları, bağımlı Saklı Markov Modelleri (HMM: Hidden Markov Model) durumlarına karşılık gelen bilgiler ile indekslenmektedir [36]. Genel olarak bir HMM yapısını temsil edilmekte ve sadece bir yoğunluk topluluğu (GMM) oluşturulmaktadır. HMM durumları ile akustik model yapısı arasında bir eşleme sağlanmaktadır. Kaldi'de konuşmacı adaptasyonu ve maksimum olasılıklı doğrusal dönüşüm için yarı bağlı kovaryans ve lineer dönüşümler uygulanmaktadır.

Dil modelleme, dildeki kelimelerin dizilişini akustik özelliklerinden tamamen bağımsız olarak modellemektedir [5]. Genel olarak n-gram tabanlı modelleme yöntemi kullanmaktadır. Bu modeller de Markov varsayımı bulunup genelde 2 ile 4 arası geçmişteki kelime sırası göz önünde bulundurarak olasılık hesaplamalar gerçekleştirilmektedir [37]. Dolayısıyla uzun bir cümledeki kelimelerin dizilişini modellemek n-gram'lar ile mümkün olmayıp sadece kısıtlı kelime geçmişi modellenebilmektedir. İleri beslemeli sinir ağını kullanan dil modellerinde Markov varsayımı bulunmadığı için bu modeller ile kelimelerdeki uzun bağımlılıklar modellenebilmektedir [38].

Dil Modeli, bir dildeki kelimelerin ve cümlelerin yapısı ve sırasını modelleyerek o dile ait bir istatistiksel model üretmektedir [39]. En basit ifade ile dil modeli bir kelime dizisinden sonra hangi kelimelerin gelebileceğini modelleyip kod çözme zamanında olası dizilişleri üretmektedir. Kaldi, Sonlu Durum Dönüştürücüleri (FST: Finite State Transducer) tabanlı bir çerçeve kullanmaktadır [40]. Bu nedenle prensipte FST olarak temsil edilebilecek herhangi bir dil modelini kullanmak mümkündür. Standart ARPA biçimindeki dil modellerini FST'lere dönüştürmek için gerekli materyaller Kaldi araç setinde yer almaktadır.

B. 3. Kod Çözme

Kaldi eğitim ve kod çözme işlemlerinde Ağırlıklı Sonlu Durum Transdüserlerini (WFST: Weight Finite State Transducer) kullanmaktadır. Kaldi'de bulunan giriş sembolleri sayısal olarak tutulmaktadır. Ancak, farklı fonemlerin aynı bilgiyi paylaşmasına izin verildiği için bu yaklaşımda bazı sorunlar ortaya çıkmaktadır. Bunlardan ilki FST'leri belirleyememek diğeri ise bir FST üzerinden Viterbi işlemi için yeterli bilgiye sahip olunamamasıdır. Viterbi kod çözme algoritması, katlamalı kodlar için en çok uygulanan kod çözme algoritmasıdır ve en büyük olabilirlik (ML: Maximum Likelihood) kod çözme tekniğini kullanmaktadır [41]. Basit FST grafiklerinden oluşturulan kod çözme grafiği HCLG olarak gerçekleştirilir. HCLG Denklemi 3'te gösterilmiştir.

$$HCLG = H \circ C \circ L \circ G \tag{3}$$

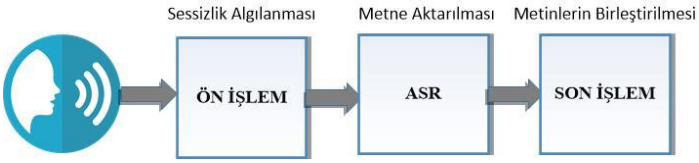
Denklem 3'te gösterilen ° sembolü, FST üzerindeki bileşimin ikili çalışmasını temsil etmektedir. G sembolü, dilbilgisi veya dil modelini kodlayan bir alıcı birimi temsil etmektedir. L sembolü, sözlüğü temsil etmektedir. C sembolü ise girişte birbirine bağımlı fonemler ve çıkıştaki fonemler arasındaki ilişkiyi temsil etmektedir. Sisteme giriş sembolleri fonemler ve çıkış sembolleri kelimeler olarak tanımlanmaktadır. H giriş kimliği olarak kabul edilen ve içeriğe bağlı fonemleri döndüren HMM durumlarını temsil etmektedir. FST'lerin girişine fonem kimliğini kodlayan "geçiş kimliği" adı verilen bir tamsayı belirteci konulmuştur. Modeldeki "geçiş kimlikleri" ile geçiş olasılık parametreleri arasında bire bir eşleme oluşmaktadır. Modeldeki stokastik yaklaşımı elde etmek için eğitim sırasında ağırlıklandırma işlemi sürekli gözden geçirilmektedir. Fonemlerin elde edilmesinden sonra yerleştirilen grafiğin boyutunun yüksek olmaması için ağırlıklandırma işlemi önemlidir.

III. UYGULAMA GELİŞTİRME

Çalışmanın bu bölümünde önerilen yöntemin ve uygulamanın sistem mimarisi verilmiştir. Önerilen yöntemdeki bazı sınıf ve değişkenlerin yapısı açıklanmıştır. Yazılım kodlarının geliştirilmesi için Java ve C++ kullanılmıştır.

A. SİSTEMİN GENEL MİMARİSİ

Sistemin genel mimarisi ön işlem, ASR ve son işlem olmak üzere üç temel modülden oluşmaktadır. Bu modüller sıralı bir şekilde çalışmaktadır. Şekil 3'te sistemin genel mimarisi verilmiştir.



Şekil 3. *Sistemin genel mimarisi*

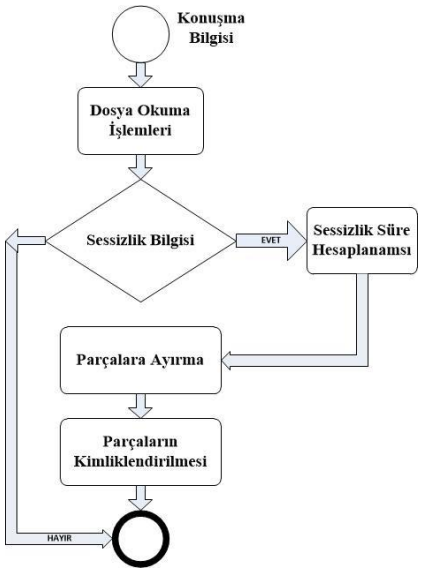
Şekil 3'te görüldüğü gibi sisteme öncelikle bir konuşma kaydı giriş olarak verilmiştir. Ön işlem modülünde konuşma içerisindeki sessizlik bilgisi elde edilmiştir. Sessizlik bilgisi ile karşılaşıldığında konuşma parçalara ayrılmıştır. Her konuşma parçası sıralı bir şekilde tek tek bir sonraki modül olan ASR modülüne verilmiştir. ASR'nin görevi kendisine gelen her konuşma parçasını metne aktarmaktır. ASR sistemi ile metne aktarılan her konuşma parçası son işlem modülünde sırasıyla birleştirilerek sunulmuştur.

B. ÖN İŞLEM MODÜLÜNÜN GELİŞTİRİLMESİ

Ön işlem modülünde öncelikle konuşma bilgisinin örneklem frekansına bakılmıştır. Materyal ve metod bölümünde detaylı olarak belirtildiği gibi örneklem frekansının belirli bir oranı alınmıştır. Böylelikle bir eşik değeri elde edilmiştir. Konuşmada içerisinde sessizlik bilgisi ile karşılaşıldığında bir sayaç işlemi uygulanmıştır. Bu sayaç sessizlik süresinin uzunluğunun hesaplanması için kullanılmıştır. Gerçekleştirilen işlemin akış şeması şekil 4'te verilmiştir.

Uygulama geliştirilirken Java'nın ön tanımlı paketlerinden biri olan "Java IO" ön tanımlı paketi dosya okuma ve parçalanmış dosyaları yazmak için kullanılmıştır [42]. Bu ön tanımlı paket altında bulunan "Java Wav File IO" sınıfı ile konuşma dosyaları üzerindeki gerekli okuma/yazma işlemleri gerçekleştirilmiştir. Ele alınan her dosyadaki konuşma bilgisinde sessizlik eşik değerinin altına düşüldüğü anda "silenceCounter" değişkeni kendisini konuşma bilgisi boyunca arttırmaktadır. Bu noktada "isSilence" fonksiyonumuz devreye girmekte ve herhangi bir anda eğer "silenceCounter != 0" bilgisine ulaşır ise konuşmayı parçalara ayırma işlemi yapılmamaktadır. Çalışmada sessizlik eşik değeri 1 saniye içerisinde 8 defa ölçülmüştür. Ardından bu değerlerin ortalaması alınmıştır. Eğer 1 saniyelik ölçümün ortalaması eşik değerinin altında ise parçalama işlemi devreye girmekte ve konuşma parçalara

ayrılmaktadır. Ayrılan her parça başlangıç ve bitiş değerine sahip olacak şekilde isimlendirilmiştir. Böylelikle parçaların ASR sistemine giriş sırası belirlenmiştir.



Şekil 4. *Ön işlem modülü için akış şeması*

C. ASR MODÜLÜ

Kaldi araç setinin 5.0 versiyonu kullanılarak bir ASR sistemi geliştirilmiştir. ASR sisteminin geliştirilmesinde Boğaziçi Üniversitesi tarafından 2012 yılında hazırlanan ve Dilsel Veri Konsorsiyumu (LDC: Linguistic Data Consortium) tarafından sunulan Türkçe konuşma veri seti kullanılmıştır [43]. Bu veri seti yaklaşık 91.44 saatlik gerçek konuşma verisi içermektedir. Konuşma dosyaları 16 KHz örneklem frekansına sahiptir.

Akustik modelleri eğitmek ve test etmek için kullanılan her bir konuşmacının akustik meta-verileri oluşturulmuştur. Veriler akustik veriler ve dil verileri olarak iki bölüme ayrılmıştır. Kaldi'nin gereksinim duyduğu akustik veriler için zorunlu olan meta verilerinin hazırlanması adımları verilmiştir [33]:

- spk2gender⇒〈 Konuşmacı kimliği〉 〈cinsiyet〉

Bu adımda konuşmacıların cinsiyeti hakkında bilgi verilmiştir. Konuşmacı kimliği, her konuşmacının benzersiz bir adıdır ve kayıt kimliği olarak adlandırılmaktadır.

- uiau.scp ⇒〈Konuşma kimliği〉 〈kaydet.wav konuşma dosyasının yolu〉

Kaydedilen konuşma dosyalarının yolu bu adımda belirtilmiştir.

- metin ⇒〈Konuşma kimliği〉 〈transkripsiyon〉

Bu adımda metin-konuşma veya fonem eşleştirmesi yani transkripsiyon işlemi gerçekleştirilmiştir.

- corpus.txt ⇒〈transkripsiyon〉

Akustik modeli oluşturmak için kullanılan tüm transkripsiyon bu adımda hazırlanmıştır.
Kaldi'nin gereksinim duyduğu dil verilerini hazırlamak için meta veriler aşağıda verilmiştir:

- lexicon.txt〈kelime〉 〈fonem 1〈fonem 2〉 .

Bu bildirimde her kelimenin fonem transkripsiyonu verilmiştir.

- tum.silencc jshones.t.ri »» (fonem).

Bu bildirim dil bilgisini elde edebilmek için ASR'de kullanılan tüm fonemleri içermektedir.

Belirtilen adımlar aynı zamanda Kaldi'de belirtilen dosya isimleri ve işlevleri temsil etmektedir. Bu nedenle Kaldi kullanım kolaylığı ve işlemlerin adım adım gerçekleşmesini sağlamıştır. Bu kullanım şekli herhangi bir adımda hata olması durumunda daha önceki adımların tekrarlanmamasını sağlamaktadır.

Belirtilen adımlar oluşturularak Kaldi için gerekli bir veri seti yapısı oluşturulmuştur. Ancak öznitelik çıkarım işlemleri için Mel frekans ölçeği kullanılmıştır. Mel frekans ölçeği, insan kulağının ses frekanslarındaki değişimi algılayışını gösteren bir ölçektir. 1000 Hz'e kadar olan seslerin algılaması doğrusal olmakta iken, frekans arttıkça değişimin algılanması logaritmik bir hale gelmektedir. MFCC, ses sinyalinin kısa-zamanlı güç spektrumunun Mel ölçeği üzerindeki ifadesidir. 25 ms.'lik karelere ayrılan ses dosyalarında güç spektrumlarına bakılarak özellikler çıkarılmıştır.

Özellikleri çıkarılan konuşma ve birebir eşleştirilmiş metin dosyaları kullanılarak geliştirilen ASR sistemi GMM-HMM tabanlı olarak geliştirilmiştir. Klasik GMM-HMM kullanılarak geliştirilen ASR sisteminde her HMM durumunun gözlem olasılığı $bj\ (x)$ Gauss karışımları kullanılarak hesaplanmıştır. Teknik olarak GMM-HMM sistemini eğitirken, monofon modeli ifade seviyesi transkriptlerinden yararlanılmıştır.

Geliştirdiğimiz Türkçe dil modelimiz, Türkçe internet gazetelerinden elde edilen metin veri setinden eğitilmiştir. Toplam büyüklük yaklaşık olarak 1.26 milyon kelimedir. Dil modelimiz N-gram tabanlı olarak geliştirilmiştir ve N değeri 3 olarak belirlenmiştir. Sözlük ise veri seti içerinde yer alan kelimelerden hazırlanmış ve fonem etiketleri tarafımızdan kontrol edilmiştir.

D. SON İŞLEM MODÜLÜ

Bu modülde ASR sisteminden elde edilen metin verileri ASR sistemine giren konuşma parçalarının sırasına göre birleştirilmiştir. Burada basit bit metin birleştirici uygulama yazılmıştır. Uygulama geliştirilirken "Java String" ana sınıfına ait "Java String join()" metodu kullanılmıştır [26].

IV. DENEYSEL SONUÇLAR

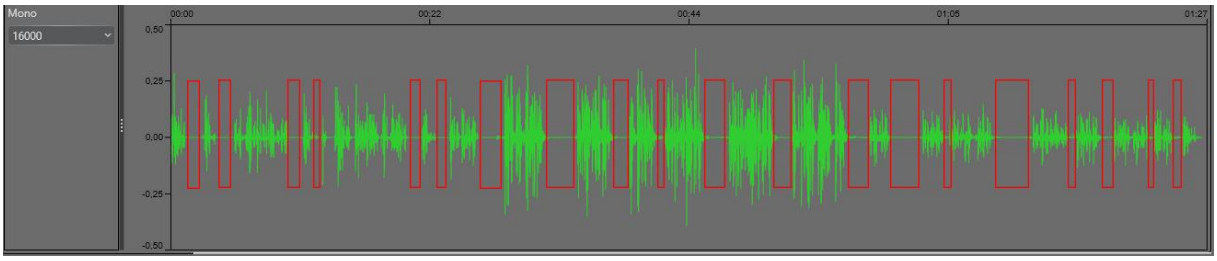
Bu çalışmada elde edilen sonuçlar farklı deneyler için Kelime Hata Oranı (WER: Word Error Rate) olarak ifade edilmiştir. WER konuşma tanıma için bilinen en güvenilir sonucu veren bir başarım hesaplama metriğidir [6]. Fonem seviyesi yerine kelime düzeyinde çalışmaktadır. ASR sisteminin çıkışında elde edilen metin verisindeki kelimeler tek tek ele alınmıştır. Her kelime olması gerektiği hali ile karşılaştırılmıştır. Herhangi bir farklılık bulunan kelime hatalı olarak nitelendirilmiştir.

Çalışmada temelde iki yaklaşım karşılaştırılmıştır. Bu yaklaşımlardan ilki klasik GMM-HMM tabanlı ASR sistemidir. Diğeri ise çalışma kapsamında önerilen yöntem olan sessizlik bilgisinin ortadan kaldırıldığı ve konuşmanın parçalara ayrıldığı yaklaşımdır. Bu yaklaşımda bilinmesi gereken diğer bir konu ASR sistemine daha küçük konuşma parçalarının verildiğidir. Klasik GMM-HMM tabanlı ASR sisteminin başarımı Tablo 1'de verilmiştir.

Tablo 1. *GMM-HMM tabanlı Türkçe ASR sistemi*

Eğitim veri seti	Test veri seti	WER(%)	Örneklem frekansı
Boğaziçi veri seti	Boğaziçi veri seti	27.70	16 KHz

Tablo 1'de ASR sisteminin geliştirilmesinde ve test edilmesinde kullanılan veri seti bilgileri verilmiştir. Eğitim ve test işlemlerinde veri setinin 1/3'ü test amaçlı ayrılmıştır. Test işlemleri Boğaziçi veri setinin test işlemleri için ayrılan kısmı kullanılmıştır. Ancak bu verilerdeki konuşmalarda yeteri kadar sessizlik veya durağan konuşmaya rastlanılmamıştır. Bu nedenle Şekil 5'te zaman-genlik dalga formu verilen sessizlik içeren Türkçe konuşma dosyaları oluşturulmuştur.



Şekil 5. *Örnek konuşma zaman-genlik dalga formu*

Şekil 5'te gösterilen zaman-genlik dalga formunda sessizlik olan yerler net bir şekilde belirtilmiştir. Şekilde gösterilen özelliğe sahip bir konuşma dosyasında gözle görülebilir 21 adet konuşma parçasının ayrı ayrı ASR'ye verilmesi gerekir. Ancak yapılan deneylerde 22 adet konuşma parçası ASR'ye giriş olarak verilmiştir. Gözle görülmeyen özellikteki sessizlik bilgisi de geliştirdiğimiz sistem tarafından algılanmıştır. Test işlemleri WER temel alınarak değerlendirildiğinde tablo 2'deki sonuçlar elde edilmiştir.

Tablo 2. *Geliştirilen uygulamanın WER sonuçları*

Uygulama yöntemi	WER(%)	Örneklem frekansı
Standart ASR	32.77	16 KHz
Ön işlemli ASR	25.36	16 KHz

Tablo 2'de belirtilen uygulama yönteminde iki farklı yöntem test edilmiştir. Standart ASR yönteminde konuşma bilgisi ASR sistemine direk verilmiş herhangi bir ön işlem yapılmamıştır. Ön işlemli ASR yönteminde ise sessizlik bilgisi ortadan kaldırılmış ve konuşma parçalara ayrılmıştır. Yapılan testlerde çalışmada sunulan ön işlemli ASR'nin kelime hata oranı açısından daha iyi sonuç verdiği görülmüştür.

V. SONUÇ VE ÖNERİLER

ASR sistemlerinde ön işlemlerin yapılması ve ASR sistemlerinin başarımını kötü etkileyecek akustik bilgilerin ortadan kaldırılması amacıyla gerçekleştirilen bu çalışmada konuşma içerisindeki sessizlik alanları kaldırılmıştır. Ayrıca akustik bilgide uzun bağımlılıklar oluşturmamak adına konuşma bilgisi parçalara ayrılmıştır. Konuşma bilgisini metne aktarabilmek için temel bir Türkçe ASR sistemi geliştirilmiştir. Parçalanan her konuşma bilgisi ayrı olarak ASR sistemine verilmiştir. ASR sisteminin çıkışında konuşma parçalarına karşılık gelen metin çıktıları birleştirilerek sunulmuştur.

Gerçekleştirdiğimiz deneylerde sessizliğin kaldırılmasının ASR sistemlerinin başarımını doğrudan etkilediği görülmüştür. Ancak bazı durumlarda konuşmanın parçalara ayrılmasının ASR sisteminin başarımının kötü etkilendiği gözlemlenmiştir. Bunun nedeni ise fonem bilgisinin tam olarak çıkarılamayacağı kadar küçük konuşma parçalarının ASR sistemine verilmesidir. Ayrıca bazı durağan konuşmacıların heceler arasında sessiz kalmaları sadece hecelerin parçalara ayrılmasına neden olmuştur. ASR sistemi ile metne aktarılan heceler son işlem modülünde birleştirilirken bazı sorunlar ile karşılaştırılmıştır. Bu sorunların başında heceler arasındaki boşluklar gelmektedir. Kelime içi boşluklar kelime hata oranının artmasına neden olmuştur. Bu nedenle ön işlem modülünde konuşmacıların yapısal özellikleri yani durağan, kekeme veya çok hızlı konuşan konuşmacıların tespit edilmesi önemlidir.

Gelecek çalışmalarda tespit edilen konuşmacı özelliklerine uygun ön işlemelerin yapılması ASR sisteminin başarımını arttıracağı düşünülmektedir. Ayrıca sadece sessizlik bilgisi değil farklı gürültü akustik bilgilerinin de konuşmadan çıkarılması çalışılabilir. Diğer yandan minimum parça uzunluğunun belirlenmesi, sadece GMM-HMM tabanlı ASR sistemlerinde değil derin sinir ağı tabanlı yaklaşımlarda da önerilen yöntemin test edilmesi yapılabilecek çalışmalar arasındadır.

TEŞEKKÜR: Bu çalışma, EMFA Yazılım Danışmanlık A.Ş. tarafından desteklenmiştir. Desteklerinden dolayı EMFA Yazılım Danışmanlık A.Ş. yönetim kurulu başkanı Emre EVREN teşekkür ederiz.

VI. KAYNAKLAR

[1] M. Abushariah, S. Gunawan, O. Khalifa, ve M. Abushariah, "English digits speech recognition system based on Hidden Markov Models," International Conference on Computer and Communication Engineering, Kuala Lumpur, Malaysia, 2010, ss. 1–5.

[2] H. Prakoso, R. Ferdiana, ve R. Hartanto, "Indonesian Automatic Speech Recognition system using CMUSphinx toolkit and limited dataset," International Symposium on Electronics and Smart Devices, Bandung, Indonesia, 2016, ss. 283–286.

[3] C. Kurian, ve K. Balakrishnan, "Speech recognition of Malayalam numbers," World Congress Natural Biology Inspired Compututer, Coimbatore, India, 2009, ss. 1475–1479.

[4] C. Howard, ve D. David, "Automatic Measurement of Speech Recognition Performance: A Comparison of Six Speaker-Dependent Recognition Devices," *Computer Speech & Language*, c. 2, s. 2, ss. 87-108, 1987.

[5] D. Amodei, "Deep speech 2:end-to-end speech recognition in english and mandarin," International Conference on International Conference on Machine Learning, New York, USA, 2006, ss. 1–28.

[6] Y. G. Thimmaraja ve H. S. Jayanna, "Creating language and acoustic models using Kaldi to build an automatic speech recognition system for Kannada language," International Conference on Recent Trends in Electronics, Information & Communication Technology, Bangalore, India, 2017, ss. 161–165.

[7] E. Bocchieri, ve D. Caseiro, "Use of geographical meta-data in ASR language and acoustic models," International Conference on Acoustics, Speech and Signal Processing, Dallas, TX, USA, 2010, ss. 5118–5121.

[8] J. Neto, "Speaker-adaptation for hybrid HMM-ANN continuous speech recognition system," European Conference on Speech Communication and Technology, Madrid, Spain, 1995, ss. 2171–2174.

[9] G. Hinton, "Deep Neural Networks for Acoustic Modeling in Speech Recognition," *Signal Processing Magazine*, c. 29, s. 6, ss. 82–97, 2012.

[10] W. Chan, ve I. Lane, "Deep convolutional neural networks for acoustic modeling in low resource languages," International Conference on Acoustics, Speech and Signal Processing, Brisbane, QLD, Australia, 2015, ss. 2056–2060.

[11] C. Ris, ve S. Dupont, "Assessing local noise level estimation methods: Application to noise robust ASR," *Speech Communication*, c. 34, s. 1, ss. 141-158, 2001.

[12] C. Guoguo, X. Hainan, W. Minhua, P. Daniel, ve K. Sanjeev, "Pronunciation and silence probability modeling for ASR," Annual Conference of the International Speech Communication Association, Dresden, Germany, 2015, ss. 533-537.

[13] L. Karray, ve A. Martin, "Toward improving speech detection robustness for speech recognition in adverse environments," *Speech Communication*, c. 40, s. 3, ss. 261–276, 2003.

[14] J. Ramírez, J.C. Segura, C. Benítez, ve A. Torre, "A new adaptive longterm spectral estimation voice activity detector," European Conference on Speech Communication and Technology, Geneva, Switzerland, 2003, ss. 3041–3044.

[15] J. Ramírez, "Spectral estimation voice activity detector," European Conference on Speech Communication and Technology, Geneva, Switzerland, 2003, ss. 3121–3125.

[16] ITU-T Recommendation G.729-Annex B. "*A silence compression scheme for G.729 optimized for terminals conforming to recommendatio*n," c. 70, 1996.

[17] F. Basbug, K. Swaminathan, ve S. Nandkumar, "Noise reduction and echo cancellation front-end for speech codecs," *Transaction Speech Audio Processing*, c. 11, s. 1, ss. 1–13, 2004.

[18] S. Gustafsson, R. Martin, P. Jax, ve P. Vary, "A psychoacoustic approach to combined acoustic echo cancellation and noise reduction," *Transaction Speech and Audio Processing*, c. 10, s. 5, ss. 245–256, 2002.

[19] J. Sohn, N.S. Kim, ve W. Sung, "A statistical model-based voice activity detection," *Signal Processing Letters*, c. 16, s. 1, ss. 1–3, 1999.

[20] S. Gazor, ve W. Zhang, "A soft voice activity detector based on a Laplacian-Gaussian model," *Transaction Speech Audio Processing*, c. 11, s. 5, ss. 498–505, 2003.

[21] L. Armani, M. Matassoni, M. Omologo, ve P. Svaizer, "Use of a CSP-based voice activity detector for distant-talking ASR," European Conference on Speech Communication and Technology, Geneva, Switzerland, 2003, ss. 501–504.

[22] K. Woo, T. Yang, K. Park, ve C. Lee, "Robust voice activity detection algorithm for estimating noise spectrum," *Electronics Letters*, c. 36, s. 2, ss. 180–181, 2000.

[23] M. Marzinzik, ve B. Kollmeier, "Speech pause detection for noise spectrum estimation by tracking power envelope dynamics," *Transaction Speech Audio Processing*, c. 10, s. 6, ss. 341–351, 2002.

[24] R. Chengalvarayan, "Robust energy normalization using speech/non-speech discriminator for German connected digit recognition," European Conference on Speech Communication and Technology, Budapest, Hungary, 1999, ss. 61–64.

[25] M. Marzinzik, ve B. Kollmeier, "Speech pause detection for noise spectrum estimation by tracking power envelope dynamics," *Transaction Speech Audio Processing*, c. 10, s. 6, ss. 341–351, 2002.

[26] J. Zheng, Q. Zhou, ve C. Lee, "Robust, real-time endpoint detector with energy normalization for ASR in adverse environments," International Conference on Acoustics, Speech, and Signal Processing, Lake City, UT, USA, 2001, ss. 233-236.

[27] C. Suyanto, "Signal energy-based automatic speech splitter: A tool for developing speech corpus," Region 10 Conference, Taipei, Taiwan, 2007, ss. 2–5.

[28] M. Asadullah, ve S. Nisar, "A silence removal and endpoint detection approach for speech processing," 3rd International Multidisciplinary Research Conference On Global Prosperity through Research & Innovation, Peşaver, Pakistan, 2013, ss. 10-15.

[29] X. Huang, ve L. Deng, "An overview of Modern Speech Recognition," *Handbook Natural Language Processing*, 1. baskı, London, England: Chapman and Hall, 2010, böl. 3, ss. 339–367.

[30] D. Povey et al., "The Kaldi speech recognition toolkit," Transactions on Audio, Speech, and Language Processing, Workshop on Automatic Speech Recognition and Understanding, Olomouc, Czech Republic, 2014, ss.1–4.

[31] S. Narang, ve M. Divya Gupta, "Speech Feature Extraction Techniques: A Review," International Journal of Computer Science and Mobile Computing, c. 4, s. 3, ss. 107–114, 2015.

[32] A. Guglani, ve N. Mishra, "Continuous Punjabi Speech Recognition Model Based on Kaldi ASR Toolkit," *International Journal of Speech Technology,* c. 18, s. 3, ss.1–6, 2018.

[33] B. Tombaloğlu, ve H. Erdem, "Development of a MFCC-SVM based Turkish speech recognition system," Signal Processing and Communication Application Conference, Zonguldak, Türkiye, 2016, ss. 1–4.

[34] A. R. Yuliani, R. Sustika, R. S. Yuwana, ve H. F. Pardede, "Feature transformations for robust speech recognition in reverberant conditions," International Conference on Computer, Control, Informatics and its Applications, Jakarta, Indonesia, 2017, ss. 57-62.

[35] A. V. Haridas, R. Marimuthu, ve V. G. Sivakumar, "A Critical Review and Analysis on Techniques of Speech Recognition: The Road Ahead," *International Journal of Knowledge-Based and Intelligent Engineering Systems*, c. 22, s. 1, ss. 39–57, 2018.

[36] M. Shahin, B. Ahmed, J. Mckechnie, K. Ballard, ve R. Gutierrez-osuna, "A comparison of GMM-HMM and DNN-HMM based pronunciation verification techniques for use in the assessment of childhood apraxia of speech," Annual Conference of the International Speech Communication Association, Singapore, Singapore, 2014, ss.1583-1590.

[37] L. Saul, ve F. Pereira, "Aggregate and mixed-order Markov models for statistical language processing," International Conference on Empirical Methods in Natural Language Processing, New Jersey, USA, 1997, ss.81-19.

[38] N. Guglani, ve J. Mishra, "Continuous Punjabi Speech Recognition Model Based on Kaldi ASR Toolkit," *International Journal Speech Technology*, c. 17, s. 1, ss. 1–6, 2018.

[39] N. John, J. Wendy, ve N. Philip, "Sing formant frequencies in speech recognition," 5th European Conference on Speech Communication and Technology, Rhodes, Greece, 1997, ss. 22-28.

[40] S. Chowdhury, U. Garain, ve T. Chattopadhyay, "A Weighted Finite-State Transducer (WFST)-based language model for online Indic script handwriting recognition, " International Conference on Document Analysis and Recognition, Beijing, China, 2011, ss. 599–602.

[41] V. Shah, R. Anstotz, I. Obeid, ve J. Picone, "Adapting an ASR to event classification of electroencephalograms," *Signal Processing Medical Biology*, Pennsylvania, USA, 2018, ss. 1–5.

[42] P. Chan, ve R. Lee, *The Java class libraries : an annotated reference,* 1. baskı, Boston, USA: Addison-Wesley, 1997, böl. 3, ss. 266-310.

[43] E. Arısoy, D. Can, S. Parlak, M. Saraçlar, ve H. Sak, "Turkish Broadcast News Transcription and Retrieval, " *Transactions on Audio, Speech, and Language Processing*, c. 17, s. 5, ss. 874–883, 2009.