TÜRKÇE AĞIZLARIN TANINMASINDA DERİN ÖĞRENME TEKNİĞİNİN KULLANILMASI

IDENTIFICATION OF TURKISH DIALECTS USING DEEP LEARNING TECHNIQUES

GÜLTEKİN IŞIK

DOÇ. DR. HARUN ARTUNER

Tez Danışmanı

Hacettepe Üniversitesi

Lisansüstü Eğitim – Öğretim ve Sınav Yönetmeliğinin

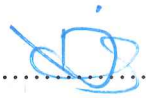
Bilgisayar Mühendisliği Anabilim Dalı için Öngördüğü

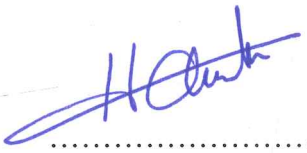
DOKTORA TEZİ

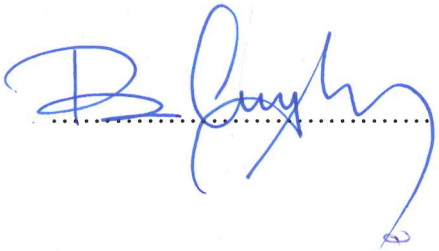
olarak hazırlanmıştır.

2019

GÜLTEKİN IŞIK'ın hazırladığı "**Türkçe Ağızların Tanınmasında Derin Öğrenme Tekniğinin Kullanılması**" adlı bu çalışma aşağıdaki jüri tarafından **BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI**'nda **DOKTORA TEZİ** olarak kabul edilmiştir.

Prof. Dr. Veysi İŞLER
Başkan ..... 

Doç. Dr. Harun ARTUNER
Danışman ..... 

Doç. Dr. Bünyamin CİYLAN
Üye ..... 

Doç. Dr. Kayhan M. İMRE
Üye ..... 

Dr. Öğr. Üyesi Burcu CAN BUĞLALILAR
Üye ..... 

Bu tez Hacettepe Üniversitesi Fen Bilimleri Enstitüsü tarafından **DOKTORA TEZİ** olarak …./...../..... tarihinde onaylanmıştır.

Prof. Dr. Menemşe GÜMÜŞDERELİOĞLU
Fen Bilimleri Enstitüsü Müdürü

Güneş'e

ve Aydede'ye…

ETİK

Hacettepe Üniversitesi Fen Bilimleri Enstitüsü, tez yazım kurallarına uygun olarak hazırladığım bu tez çalışmasında,

- tez içindeki bütün bilgi ve belgeleri akademik kurallar çerçevesinde elde ettiğimi,
- görsel, işitsel ve yazılı tüm bilgi ve sonuçları bilimsel ahlak kurallarına uygun olarak sunduğumu,
- başkalarının eserlerinden yararlanılması durumunda ilgili eserlere bilimsel normlara uygun olarak atıfta bulunduğumu,
- atıfta bulunduğum eserlerin tümünü kaynak olarak gösterdiğimi,
- kullanılan verilerde herhangi bir tahrifat yapmadığımı,
- ve bu tezin herhangi bir bölümünü bu üniversite veya başka bir üniversitede başka bir tez çalışması olarak sunmadığımı

beyan ederim.

04 / 02 / 2019

GÜLTEKİN IŞIK

YAYIMLAMA VE FİKRİ MÜLKİYET HAKLARI BEYANI

Enstitü tarafından onaylanan lisansüstü tezimin/raporumun tamamını veya herhangi bir kısmını, basılı (kâğıt) ve elektronik formatta arşivleme ve aşağıda verilen koşullarla kullanıma açma iznini Hacettepe Üniversitesi'ne verdiğimi bildiririm. Bu izinle Üniversiteye verilen kullanım hakları dışındaki tüm fikri mülkiyet haklarım bende kalacak, tezimin tamamının ya da bir bölümünün gelecekteki çalışmalarda (makale, kitap, lisans ve patent vb.) kullanım hakları bana ait olacaktır.

Tezin kendi orijinal çalışmam olduğunu, başkalarının haklarını ihlal etmediğimi ve tezimin tek yetkili sahibi olduğumu beyan ve taahhüt ederim. Tezimde yer alan telif hakkı bulunan ve sahiplerinden yazılı izin alınarak kullanması zorunlu metinlerin yazılı izin alarak kullandığımı ve istenildiğinde suretlerini Üniversiteye teslim etmeyi taahhüt ederim.

Yükseköğretim Kurulu tarafından yayınlanan "***Lisansüstü Tezlerin Elektronik Ortamda Toplanması, Düzenlenmesi ve Erişime Açılmasına İlişkin Yönerge***" kapsamında tezim aşağıda belirtilen koşullar haricinde YÖK Ulusal Tez Merkezi / H. Ü. Kütüphaneleri Açık Erişim Sisteminde erişime açılır.

☐ Enstitü / Fakülte yönetim kurulu kararı ile tezimin erişime açılması mezuniyet tarihimden itibaren 2 yıl ertelenmiştir.

☐ Enstitü / Fakülte yönetim kurulu gerekçeli kararı ile tezimin erişime açılması mezuniyet tarihimden itibaren … ay ertelenmiştir.

☐ Tezim ile ilgili gizlilik kararı verilmiştir.

04 / 02 / 2019

GÜLTEKİN IŞIK

ÖZET

TÜRKÇE AĞIZLARIN TANINMASINDA DERİN ÖĞRENME TEKNİĞİNİN KULLANILMASI

Gültekin IŞIK

Doktora, Bilgisayar Mühendisliği Bölümü
Tez Danışmanı: Doç. Dr. Harun ARTUNER
Şubat 2019, 115 sayfa

Otomatik konuşma tanıma sistemleri, konuşma seslerinin metne geçirilmesine yarar. Herhangi bir dilde otomatik konuşma tanıma sisteminin performansı, konuşmacı cinsiyeti ve duygu durumunun yanı sıra dilin varyantları olan ağızlara da bağlıdır. Ağızlar aynı coğrafik bölgede yaşayan insanların konuştuğu, söyleyiş biçimi ve sözcüksel yapı olarak birbirine benzeyen ve diğer bölgelerde konuşulan ağızlardan, sayılan bu özellikler bakımından ayrılan günlük konuşma biçimleridir. Ağız tanımanın amacı sesli ifade özelliklerinden insanların ağızlarının belirlenmesidir. Ağzın tanınmasının ardından dil ve akustik modellerin bu ağza adapte edilmesiyle konuşma tanıma sisteminin başarımının arttığı bilinmektedir. Ayrıca sesli ifadeden konuşulan ağzın belirlenmesi; sesli yanıt sistemlerinde ön işlem adımı olarak veya adli bilişimde ipucu elde etmede kullanılabilir.

Ağız tanımada kullanılan modelleme teknikleri farklı dil katmanlarındaki bilgiyi modellemeye yöneliktir. Akustik, fonotaktik ve prozodik katmanlarındaki öznitelikler insanların konuştuğu ağza özgü önemli bilgiler vermektedir. Konuşmanın fonetik farklılıkları, fiziksel düzeyde spektral öznitelikleri incelenerek tespit edilebilmektedir. Klasikleşmiş Mel Frekans Kepstral Katsayıları (MFCC) ve Log mel-spektrogram gibi öznitelikler bu amaçla kullanılmaktadır. Fonotaktik, bir dilde/ağızda, fonemlerin bir arada

bulunma kurallarına karşılık gelmektedir. Fonem dizilimleri ve bu dizilimin sıklığı ağızdan ağza değişiklik göstermektedir. Fonem dizilimleri fonem tanıyıcılar yardımıyla elde edilmekte ve daha sonra dil modelleriyle fonem dağılımları çıkartılmaktadır. Prozodi, konuşmanın tonlama, vurgu ve ritim gibi işitsel özniteliklerdir. Bu özniteliklerin insanın konuşmayı algılamasında anahtar rol üstlendiği bilinmektedir. Bu algısal öznitelikler fiziksel düzeyde temel frekans (perde), enerji ve sürenin ölçülmesiyle elde edilmekte ve uygun parametrik gösterimlere çevrilmektedir.

Son yıllarda, derin sinir ağlarının popüler hale gelmesiyle birlikte Konvolüsyonel Sinir Ağları (CNN) özellikle görüntü ve konuşma tanımada sıklıkla kullanılmaktadır. Bunun yanı sıra Uzun Kısa-Dönem Bellekli (LSTM) yinelemeli sinir ağları dizi sınıflandırma ve dil modelleme problemlerinde çokça kullanılmaktadır. LSTM sinir ağları, uzun dönemli bağlam bilgisini modellemede n-gram modellerden daha başarılıdır.

Türkiye'nin farklı yörelerinde yaşayan insanların konuştuğu ağızlar yukarıda bahsedilen özellikler açısından birbirinden ayrılmaktadır. Bu bakımdan, bu tez çalışmasında akustik, fonotaktik ve prozodik öznitelikler kullanılarak Türkçenin ağızlarının CNN ve LSTM sinir ağlarıyla sınıflandırılması konu edilmiştir. Bu amaçla Ankara, Alanya, Kıbrıs ve Trabzon ağızlarından oluşan bir Türkçe veri kümesi oluşturulmuştur. Önerilen yöntemler Türkçe veri kümesi üzerinde sınanmış ve yorumlanmıştır. Çalışma sonucunda, kullanılan yöntemlerin Türkçe ağız tanıma için oldukça iyi sonuçlar verdiği gözlenmiştir.

Anahtar Kelimeler: Türkçe Ağız Tanıma, Konvolüsyonel Sinir Ağları, Uzun Kısa-Dönem Bellekli Yinelemeli Sinir Ağları, Akustik, Fonotaktik, Prozodi

ABSTRACT

IDENTIFICATION OF TURKISH DIALECTS USING DEEP LEARNING TECHNIQUES

Gültekin IŞIK

Doctor of Philosophy, Department of Computer Engineering
Supervisor: Assoc. Prof. Dr. Harun ARTUNER
February 2019, 115 pages

Automatic speech recognition systems are used to translate speech sounds into text. The performance of the automatic speech recognition system in any language is dependent on the speaker gender and emotion as well as dialects that are variants of the language. Dialects are the speech forms that are similar to each other in the same geographic region as the utterance and lexical structure. With these characteristics, dialects are separated from each other. The aim of the dialect recognition is to identify the humans' dialect from their speech. Following the recognition of the dialect, it is known that the performance of the speech recognition system is enhanced by adapting the language and acoustic models to this dialect. Furthermore, identifying spoken dialect from speech can be used as a preprocessing step in voice response systems, or it can help to obtain a clue in forensics.

The modeling techniques used in dialect recognition are intended to model information in different language layers. Features in the acoustics, phonotactic and prosodic layers give important information that specific to the dialect. Phonetic differences of speech can be determined by examining their spectral features at the physical level. Features such as classical Mel Frequency Cepstral Coefficients (MFCC) and Log mel-spectrogram are used for this purpose. Phonotactic corresponds to the rules of coexistence of phonemes in a

language/dialect. Phoneme sequences and the frequency of this sequence vary from dialect to dialect. Phoneme sequences are obtained by phoneme recognizers and then phoneme distributions are extracted using language models. Prosody is the auditory features of speech such as intonation, stress and rhythm. It is known that these features play a key role in the human perception of speech. These perceptual features are extracted by measuring the fundamental frequency (pitch), energy and duration at the physical level and converted into appropriate parametric representations.

In recent years, Convolutional Neural Networks (CNNs) have been frequently used particularly in image and speech recognition since deep neural networks become popular. In addition, Long Short-Term Memory (LSTM) recurrent neural networks are widely used in sequence classification and language modeling problems. LSTM neural networks are more successful in modeling long-term context information than n-gram models.

Dialects spoken by people living in different regions of Turkey are separated from each other in terms of features mentioned above. From this perspective, in this thesis, acoustics, phonotactic and prosodic features were used to classify Turkish dialects with CNN and LSTM neural networks. For this purpose, a Turkish data set consisting of Ankara, Alanya, Kıbrıs and Trabzon dialects was formed. The proposed methods have been tested and interpreted on the Turkish data set. As a result of the study, it was observed that the methods used gave very good results for Turkish dialect recognition.

Keywords: Turkish Dialect Recognition, Convolutional Neural Networks, Long Short-Term Memory Recurrent Neural Networks, Acoustics, Phonotactics, Prosody

TEŞEKKÜR

Tez konusunun belirlenmesini sağlayan, tez metninin yazılmasına ve tez çalışmasının hazırlanmasına yardımcı olan, her zaman ama özellikle doktora sürecimin en zor dönemlerinde hoşgörü ve sabırla desteğini hiç eksik etmeyen, değerli hocam, tez danışmanım Sayın Doç. Dr. Harun ARTUNER'e,

Tez izleme aşamasında çok değerli yorum ve önerileriyle tezin gelişmesine ve tez metnini inceleyerek biçim ve içerik bakımından son halini almasına yardımcı olan, nezaketleriyle örnek aldığım tez izleme komitesi üyesi hocalarım Sayın Prof. Dr. Veysi İŞLER ve Sayın Dr. Öğr. Üyesi Burcu CAN BUĞLALILAR'a,

Tez metnini inceleyerek içerik bakımından son halini almasına yardımcı olan, tez savunma sınavım sırasında önerileriyle bana katkıda bulunan hocalarım Sayın Doç. Dr. Bünyamin CİYLAN ve Sayın Doç. Dr. Kayhan M. İMRE'ye,

Eğitim ve çalışma hayatım boyunca hep gurur duyduğum ve kendimi bir parçası olarak gördüğüm Hacettepe Üniversitesi Bilgisayar Mühendisliği Bölümü çalışanlarına ve tüm değerli arkadaşlarıma,

Mensubu olmaktan kıvanç duyduğum Iğdır Üniversitesi Bilgisayar Mühendisliği Bölümü'ne,

Bu süreçte ve hep yanımda olan, beni destekleyen, büyük sevgisini ve anlayışını hiçbir zaman benden esirgemeyen hayat arkadaşım Gülşah'a,

Hayatımın her aşamasında hep yanımda olan, sevgi ve desteklerini hiç eksik etmeyen çok sevgili annem ve babama,

Canı gönülden teşekkür ederim.

İÇİNDEKİLER

ŞEKİLLER

ÇİZELGELER

SİMGELER VE KISALTMALAR

Simgeler

$P(y\|x)$	x girişine göre y'nin olasılığı
$\propto$	Öğrenme oranı
σ^2	Varyans
s	Adım boyu
$\phi()$	Softmax fonksiyonu
T	Zaman adımı
PP	Karışıklık
F_0	Temel Frekans

Kısaltmalar

CNN	Konvolüsyonel Sinir Ağları
RNN	Yinelemeli Sinir Ağları
LSTM RNN	Uzun Kısa-Dönem Bellekli Yinelemeli Sinir Ağları
PPRLM	Paralel Fonem Tanıma ve Dil Modelleme
DNN	Derin Sinir Ağları
MLP	Çok Katmanlı Algılayıcılar
OKT	Otomatik Konuşma Tanıma
IVR	Sesli Yanıt Sistemi
CE	Çapraz Entropi
BPTT	Zaman Boyunca Geri Yayılım
HMM	Saklı Markov Modeli
GMM	Gauss Karışım Modeli
MFCC	Mel Frekansı Kepstral Katsayıları
RMS	Karekök Ortalama

1. GİRİŞ

İnsanın, iş için makine kullanma serüveninin İngiltere'deki dokuma tezgahlarıyla başladığı kabul edilmektedir. O zamana kadar insanın eliyle ve bedeniyle yaptığı işler artık makinelere devrediliyordu. Kömür, buhar ve demirin gücü insanlığı tümüyle başka bir çağa taşıdı.

Yaklaşık 250 yılda gelinen noktada insanlar makinelerle daha sıkı bir ilişkiye girmiş durumdalar. İnsanlar, diğer insanlarla olduğu gibi, makinelerle de sözlü veya yazılı iletişim kurmaktalar. İnsan-makine etkileşimi yazının yavaş ve kurallı halinden, sözlü iletişimin daha hızlı ve daha az kural içeren yapısına doğru evrilmektedir.

İnsanlar sosyal varlıklardır. İnsan gün içinde yaşadığı veya çalıştığı çevrede yeni insanlarla karşılaşır, tanışır, münasebet kurar. Bu etkileşim sırasında insanlar, doğaları gereği, birbirlerinin konuşmasından kimi bilgiler yakalamaya çalışır. Bunlar kişinin konuşma tarzı, sosyokültürel yapısı, gerçek yaşı, etnik kökeni, memleketi gibi bilgiler olabilir. Çoğu defa elde ettikleri bu ipuçlarının, konuşulan kişiye sormak suretiyle ya da başka şekilde doğrusunu öğrenerek sağlamasını yaparlar. Eğer tahminleri doğruysa konu hakkındaki bilgileri pekişir, değilse yeni bir bilgi edinirler. Bir insan yaşamı boyunca binlerce insanla etkileştiğinden bu süreç de binlerce defa tekrar edilmektedir. Bu şekilde, sadece konuşma özelliklerinden kişiler hakkında bilgiler edinme konusunda bir tecrübe oluşur.

İnsanın zamanla deneyimle kazandığı yeteneği makineler de kazanabilmektedir. Makineler de bu yetenekleri, aynı insanda olduğu gibi, deneyimleyerek öğrenmektedir.

Sözlü iletişimin (ya da konuşmanın) temelini diller oluşturur. Konuşulan sözcüklerin bilgisayar tarafından tanınarak metne aktarılması otomatik konuşma tanıma (OKT) olarak anılmaktadır. Konuşmayı bilgisayara otomatik olarak tanıtmak söz konusu olduğunda birçok zorlukla karşılaşılır. Bunlar konuşmacı özellikleri, konuşmanın hızı ve konuşmanın gürültüye dayalı değişkenlikleri olabileceği gibi sözcük sayısının artması, sözcüklerin ayrık ya da sürekli olması gibi nedenler de olabilir.

Basit anlamda otomatik konuşma tanıma, her bir sözcüğü sayısal dalga formunda ve karşılık gelen seslendirmesiyle birlikte tutulan bir sözlük olarak düşünülebilir. Bir girdi

verildiğinde, sistem sözlükte bu girdinin eşleniğini ve karşılık gelen metnini bulmaktadır. Bu yaklaşım konuşmacıya bağımlı az sayıda ve ayrık sözcüğün bulunduğu durumda iyi çalışırken daha karmaşık sistemlerde yukarıda sayılan zorluklardan dolayı uygulanabilir olmaktan uzaklaşmaktadır.

En düşük değişkenliğe sahip konuşmacıya bağımlı tanıma sistemleri bile bir sözcüğün farklı söylenişlerinden etkilenebilir. Bu zorluğun yanı sıra konuşmacıdan bağımsız sistemlerin yaş, cinsiyet, konuşma tarzı, insanların değişik anatomik özellikleri ve ağız özelliklerini de dikkate alması gerekir.

Ağızlar, ait oldukları dilden bazı özellikler bakımından ayrılan ve ülkenin belli bir bölgesine özgü olan konuşma biçimleridir. Bu nedenle fonemlerin ve bunun sonucunda sözcüklerin söylenişinde farklar oluşmaktadır. Ağızlara özgü karakteristiklerin elde edilmesi ve bunlar kullanılarak ağızların tanınması, konuşma işleme alanında popüler konular arasındadır. Sadece akustik farkların değil aynı zamanda ağızlardaki farklı sözcüklerin de OKT sistemlerine dahil edilmesi gerekir. Bununla birlikte, ağız tanıma işlemi, farklı ağızlar üzerinde eğitilmiş OKT sisteminin bir ön adımı olarak düşünülebilir. Bu durumda, verilen ses örneği ilk önce ağız tanıma adımından geçirilerek hangi ağız örneği olduğu tespit edilir daha sonra ilgili OKT sistemine anahtarlama yapılır.

Ağız tanıma, belli bir dilde verilen bir konuşmanın akustik sinyalinden ağız yöresinin belirlenmesi olarak tanımlanabilir. Konuşma sinyalinin içinde diğer birçok bilgiyle (yaş, cinsiyet, duygu vb.) birlikte ağızla ilgili de bilgiler saklıdır. Ağızla ilgili bilgiler konuşma sinyalinin farklı düzeylerindedir. Parçalı sesbirim (segmental, fonem) düzeyinde, sesler üretilirken ses yolunun aldığı bir dizi farklı şekillerde ağza özgü bilgiler gözlenebilir. Parçalar-üstü (supra-segmental) sesbirim düzeyinde ağza özgü bilgiler, hecelerin süre örgüsünde, perde ve enerji eğrilerinin şeklinde gizlidir.

Standart dil, ülkede konuşulan bir ağzın resmi dil olarak benimsenmesiyle elde edilir. Örneğin Türkiye için İstanbul ağzı standart Türkçe olarak belirlenmiştir. Standart dil genelde resmi kurumlarda, okullarda, ulusal haber kanallarında kullanılır. Ağızlar ise daha gayri resmî, enformel yerlerde ve günlük yaşamda kullanılmaktadır. Ağızlar hem kendi arasında hem de standart dilden farklılık gösterir. Bu farklılık kimi zaman fonetik (sesler)

bakımından, kimi zaman ise fonolojik yani ses dizimi, tonlama ve vurgu bakımından kendini gösterir. Bunlara ek olarak sözcüksel (leksikal) farklılıklar da görülmektedir.

Yetişkin bir kişi Türkçenin ağızlarını, farkların hangi unsurlardan kaynaklandığını bilmese bile, birbirinden ayırt edebilir. Yeni biriyle tanıştığımızda, ağız özelliklerini yansıtıyorsa, tanıştığımız kişinin "nereli" ya da en azından hangi bölgeden olduğunu tahmin edebiliriz. Bu tez çalışmasının konusu da, konuşmanın sinyal düzeyinden çıkartılan bilgiler ile Türkçenin ağız yöresinin otomatik olarak tahmin edilmesidir.

Konuşulan dilin kimliklendirilmesi (identification) olarak açıklanan dil tanıma, ağız tanıma ile yakından ilişkilidir. Ağız tanıma dil tanımanın yaklaşımlarını kullanmakla beraber, dil tanımadan daha zordur. Çünkü ağızlar, diller kadar birbirinden ayrılmazlar. Biçimsel (morfolojik), sözcüksel (leksikal), söz dizimsel (sentaktik), fonetik ve fonolojik olarak ağızlar birbirinden ayrılsa da, bu ayrım dillerde daha belirgindir.

Konuşma tanımaya etkisinin ötesinde ağız tanımanın çeşitli avantajları da vardır. Sesli yanıt (Interactive Voice Response, IVR) sistemlerine ağız tanıma adımı entegre edilerek, bu sistemlerin başarımları arttırılabilir. İnsanların sosyokültürel geçmişi, etnik kökeni, doğduğu yer gibi bilgileri, konuştuğu ağızdan elde edilebilmektedir. Bu bakımdan otomatik ağız tanıma, adli ve güvenlik ile ilgili olaylarda kişinin arama kümesini büyük oranda düşürecektir. Bundan dolayı otomatik ağız tanıma çalışmaları, coğrafik orijinlerin tespit edilmesine yönelik çalışmalara kaynaklık edebilir. Bunun için, konuşulan bütün ağızların dahil edilmesine gerek olmadan sadece genel birkaç ağız bölgesiyle bile konuşma örneğinin en azından hangi ağza benzediği bulunabilir.

Bu tez çalışmasında, farklı yaklaşımlar ve farklı modelleme teknikleriyle akustik sinyaldeki bilgiler kullanılarak ağız tanıma işlemleri yapılmıştır. Bu açıdan; çerçeve tabanlı akustik yaklaşım, ses dizimine dayalı fonotaktik yaklaşım ve vurgu ve ritim gibi özelliklere dayalı prozodik yaklaşım kullanılarak Türkçenin Ankara, Alanya, Kıbrıs ve Trabzon ağızları üzerinde tanıma yapılmıştır. Bunlar için dil tanıma çalışmalarından esinlenilerek yeni modelleme yaklaşımları geliştirilmiştir. Derin öğrenme olarak adlandırılan derin sinir ağları kullanımı bu yaklaşımlarda önemli bir yer tutmaktadır.

Dikkat çekici derin öğrenme mimarilerinden konvolüsyonel sinir ağları (Convolutional Neural Networks, CNN) akustik çerçevelerin sınıflandırılmasında kullanılmıştır. Çerçevelerden çeşitli öznitelikler elde edilerek zamansal bağımlılığı da hesaba katacak şekilde derin sinir ağlarıyla tanıma yapılmıştır. Ayrıca fonotaktik yaklaşımların en bilinenlerinden biri olan Paralel Fonem Tanıma ve Dil Modelleme (Parallel Phoneme Recognition followed by Language Modelling, PPRLM) yöntemi, uzun kısa-dönem bellekli yinelemeli sinir ağlarıyla (Long Short-Term Memory Recurrent Neural Networks, LSTM RNN) yeniden yapılandırılmıştır. Bunlara ek olarak prozodik yaklaşımlarda tonlama, vurgu ve ritim özellikleri derin sinir ağları kullanılarak modellenmiş ve bu ağlar yardımıyla sınıflandırma yapılmıştır.

1.1. Özgün Katkı

Otomatik olarak yapılmış Türkçe ağız tanıma çalışması literatürde bulunmamaktadır. Bu bakımdan tez çalışması bu alanda ilk olma özelliği taşımaktadır.

Türkçenin 4 ağız bölgesini (Ankara, Alanya, Kıbrıs, Trabzon) içeren bir veri kümesi oluşturulmuştur. Bu veri kümesinin daha sonra bu alanda yapılacak çalışmalara kaynaklık edebileceği düşünülmektedir.

Cümle sonundaki sözcüklerin diğer sözcüklere göre ağız özelliğini daha çok yansıtan bilgiler içerdiği bilinmektedir. Bu bilgiden hareketle, cümle sonundaki kısa konuşma parçaları (sözcükler) kullanılarak ağızların akustik açıdan sınıflandırılabileceği gösterilmiştir. Bunun için logaritmik mel-spektrogram ve MFCC öznitelikleri sırasıyla CNN ve ileri beslemeli derin sinir ağlarında kullanılmıştır.

Fonotaktik yaklaşımın en bilinen yöntemlerinden birisi PPRLM'dir. Bu mimaride fonem tanıyıcılar paralel olarak çalışır ve ardından N-gram dil modelleri gelir. Bu tez çalışmasında LSTM sinir ağı fonem tabanlı eğitilerek dil modelleri oluşturulmuş ve PPRLM mimarisine entegre edilmiştir. Ağzın belirlenmesi kararı için ayrıca ileri beslemeli sinir ağı kullanılmıştır.

Prozodik yaklaşımda, perde ve enerji eğrilerinin Legendre polinom katsayılarıyla temsil edilmesi ve bu katsayıların öznitelik olarak kullanılarak LSTM sinir ağlarında çoka-bir sınıflandırılması gösterilmiştir. Ayrıca bu eğriler ayrık birimlere dönüştürülerek LSTM

ağlarıyla her bir ağzın profili çıkartılmıştır. Burada sınıflandırmanın temelini ağız profilleri (modelleri) oluşturmaktadır.

1.2. Tez Metninin Organizasyonu

Bu tez çalışması şu şekilde bölümlere ayrılmıştır: 2. Bölümde konuşma üretimi ve bazı dilbilimsel konular açıklanmıştır. Fonetik ve fonolojik tanımlar ve bunlar arasındaki ayrımlardan söz edilmiştir.

Çalışmanın 3. Bölümünde Türkçenin ağızlarının birbirileri arasındaki değişimler ve standart dilden farkları ele alınmıştır.

Çalışmanın 4. Bölümünde derin öğrenmeye temel teşkil eden sinir ağları ve özel yapılı sinir ağı mimarilerinden bahsedilmiştir. CNN ve LSTM tipindeki sinir ağlarının sınıflandırmada ve dil modellemede kullanılması anlatılmıştır.

5. Bölümün birinci kısmında dil ve ağız tanıma çalışmalarına ilişkin detaylı bilgiler ele alınmıştır. Bunun yanında literatür bilgileri ve kullanılan veri kümesiyle ilgili sayısal bilgiler yer almaktadır.

Konuşma sinyalinin çerçeve tabanlı akustik özniteliklerinin çıkartılması, bunların geleneksel çok katmanlı sinir ağları ve CNN mimarisiyle sınıflandırmada kullanılması 5. Bölümün ikinci kısmında anlatılmıştır. Ayrıca önerilen yöntemin TIMIT veri kümesi üzerindeki başarımı karşılaştırmalı olarak gösterilmiştir. Akustik yaklaşımın bulguları ve sonuçları açıklanmıştır.

5. Bölümün üçüncü kısmında fonotaktik yaklaşımın ayrıntıları verilmiş, PPRLM mimarisi ve bu mimaride konuşmayı fonemlerine ayırmaya yarayan fonem tanıyıcılardan bahsedilmiştir. LSTM tipindeki yinelemeli sinir ağlarının PPRLM mimarisinde dil modelleme amacıyla nasıl kullanıldığı anlatılmıştır. Yaklaşımın bulguları ve sonuçları bu kısımda verilmiştir.

5. Bölümün dördüncü kısmında, ağızların prozodik açıdan sınıflandırılması üzerinde durulmuştur. Legendre polinom katsayılarının öznitelik olarak kullanılması anlatılmıştır.

Örnek cümlelerin perde ve enerji eğrileri ayrık birimler olarak modellenmiş ve bunlar kullanılarak ağız profilleri çıkartılmıştır. Bulgular ve sonuçlar bu kısımda rapor edilmiştir.

6. Bölümde çalışmada kullanılan yaklaşımların genel sonuçları ve tartışmalara yer verilmiş ve gelecekte yapılması planlanan çalışmalardan bahsedilmiştir.

2. KONUŞMA ÜRETİMİ, SES VE DİL BİLGİSİ

İnsanlar arasındaki fikir veya bilgi paylaşımları dil aracılığıyla sağlanır. Bu açıdan dil bir iletişim aracıdır ve temel yapısı seslerden oluşur. Bu bölümde bu seslerin fiziksel olarak üretilmesi anlatılacak, ardından fonetik ve fonolojik özellikler üzerinde durulacaktır.

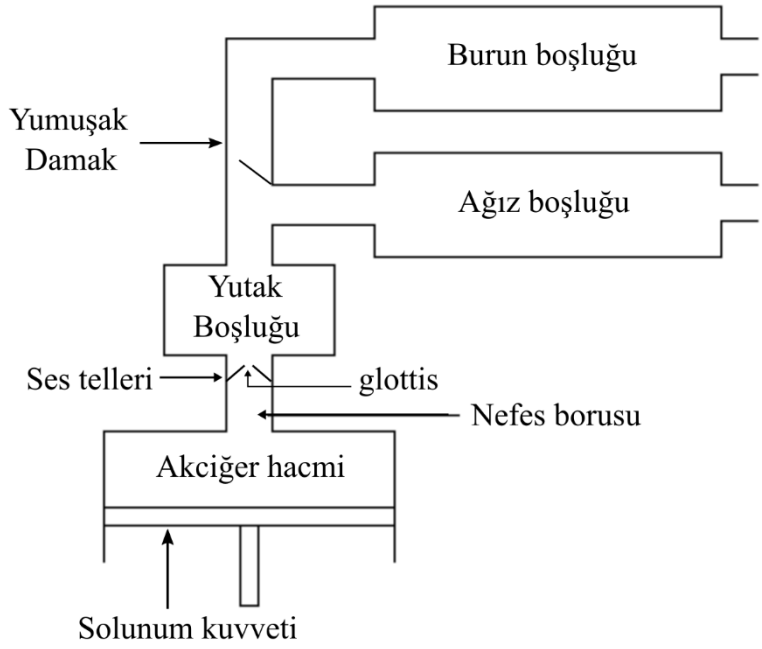
2.1. Konuşma Üretimi

İnsanın konuşma üretimi üç aşamada gerçekleşir:

1. Akciğer enerjiyi üretir (Solunum).

2. Ses telleri bu enerjiyi duyulabilir sese çevirir (Ses üretimi (fonasyon), gırtlak).

3. Ağız ve burun boşlukları sesi anlaşılır konuşmaya çevirir (söyleyiş, artikülasyon).



Şekil 2.1 Ses yolunun basitleştirilmiş diyagramı.

Diyaframın indirilip kaburgaların kaldırılmasıyla göğüs boşluğu şişer ve akciğerin hacmi arttığı için hava basıncı azalır. Böylece basıncı eşitlemek için dışarıdan hava alınır. Tam tersi süreçte ise, diyafram yukarı çekilip kaburgalar indirilerek göğüs kafesi büzülür ve akciğerlerdeki hava basıncı artar. Böylece basıncı dengelemek için hava akciğerlerden dışarı atılır.

Konuşma sesinin kaynağı, yukarıda bahsedilen, akciğerlerden dışarı atılan havadır. Solunum sırasında ses telleri hava akışını serbest bırakmak için glottis adı verilen bir boşluk oluşturur. Buradan çok az duyulabilen (veya hiç duyulamayan) bir ses çıkar.

Sesler çıkarılırken ses telleri glottis'i kapatarak akciğerlerin yüksek basınçla (alt gırtlaksı basınç) dolmasına neden olur. Bu basınç yeterince yüksek olduğunda ses telleri birbirinden ayrılmaya başlar ve hava akışının glottis'ten yutağa geçmesini sağlar. Ses telleri elastik ve aerodinamik yapısından dolayı geri yaylanır, tekrar basınç oluşur, ses telleri tekrar açılır... Ses telleri kapalı kaldıkça ve yeterli alt gırtlaksı basınç oluştukça bu süreç böyle devam eder. Ses üretimi (fonasyon) olarak bilenen bu periyodik süreç ötümlü seslerin üretilmesini sağlar. Bu periyodik süreçteki salınımın temel frekansı, ses tellerinin uzunluğu ve kütlesi ile belirlenirken ses tellerinin gerginliği ile kontrol edilir. Erkeklerde temel frekans 80-200 Hz arasındayken, kadınlarda 150-300 Hz arasındadır.

Fısıldama sesinin üretiminde ses telleri titreşmez. Üçgen boşluk oluşturacak şekilde birbirine yakın dururlar. Glottis'ten akan hava fısıldama sesine benzeyen ötümsüz seslerin çıkmasına neden olur.

Gırtlaktan üretilen ses; gırtlağın üzerindeki ses yolunda bulunan zar (yutak içine doğru burun boşluğu bağlantısını açar ve kapatır), diş, dil, dudak ve çene pozisyonları üzerinde oynanarak değişime uğratılabilir. Bu organların pozisyonuna bağlı olarak farklı tınlama frekansları oluşur. Bu tınlama frekansları formant olarak adlandırılır ve konuşmanın anlaşılırlığı için gereklidir. Tınlama frekansı, glottis'ten kaynaklanan temel frekanstan etkilenip değişmediği için, ünlüler her farklı temel frekansta tanınabilir.

İnsanın konuşma aygıtı, seslerin üretimi için yukarıda sayılan durumları kombine ederek teorik olarak sonsuz sayıda farklı ses üretebilir. Ancak her dil sınırlı sayıda sesbirime sahiptir.

2.2. Fonetik ve Fonolojik Özellikler

Bu tez çalışmasında, ağızlar hem fonetik (akustik) hem de fonolojik (fonotaktik, prozodik) özellikler bakımından sınıflandırıldığı için iki terimin de esaslarını tartışmak gerekir.

Fonetik bir dildeki tanımlanabilen, fiziksel olarak ölçülebilen konuşma seslerini inceler. Seslerin söylenişinden akustik özelliklerine ve beyinde algılanmasına kadar olan üç özelleşmiş kısma ayrılır [1]. Birinci kısım; seslerin üretilişi sırasında dudak, dil, diş, çene ve ses telleri gibi konuşma organlarının pozisyonu, hareketi ve şeklinin incelenmesi ve buna bağlı olarak seslerin sınıflandırılması ile ilgilenir. Temelde ünlü-ünsüz ve daha sonra dar, geniş, düz, yuvarlak gibi sınıflandırmalar bu kısımda yapılır. İkinci kısım konuşma ile üretilen ses dalgalarının frekans ve enerji gibi ölçülebilir fiziksel özelliklerini inceler. Bu özellikler kişiden kişiye değişebildiği gibi dilden dile veya ağızdan ağza da değişebilmektedir. Üçüncü kısım, konuşma seslerinin insanda algılanma ve tanınmasını inceler. Bazı sesler, bazı dillerde olmadığı için o seslerin algılanma biçimi de o dili konuşanlar arasında ayrıma neden olmaktadır.

Ünlü sesler ve ünsüzlerin bazıları (/b, c, d, g, j, l, m, n, r, v, y, z/) gırtlakta üretilirken titreşime neden olduklarından bu seslere ötümlü (voiced) sesler denir. Ötümlü seslerin sinyali periyodiktir. Yani genlik-zaman gösteriminde birbirini tekrar eden bir örüntüye sahiptir. Gırtlakta titreşime neden olmayan bir kısım ünsüz (/f, s, t, k, ç, ş, h, p/) ise ötümsüz (voiceless, unvoiced) seslerdir. Bunların sinyali genlik-zaman gösteriminde aperiyodik özelliktedir. Bu ses sinyallerinin herhangi bir örüntüsü yoktur ve rastgeledir. Bir sesin ötümlü veya ötümsüz olduğu, ses çıkarılırken elle gırtlağa dokunarak hissedilebilir. Ancak daha kesin yolu spektrogramlar aracılığıyla akustik analiz ile mümkündür. Ötümlü sesler düşük frekanslarda yüksek enerjili bölgelerde görülür [2].

Genlik-zaman gösteriminde sinyalin periyodik olup olmamasına bakılarak çıkarılan sesin ötümlü veya ötümsüz olduğu tespit edilebilir. Ancak bu gösterimden sesbirimlerin (fonemler) kimliği belirlenemez. Bunun için dalga formundan frekans alanına geçiş yapmak gerekir.

Fonoloji, belli bir dildeki seslerin işlevleri ve diziliş kuralları gibi daha soyut kavramları inceler [3]. Sesler, anlam ayıran sesler, sesdizimi, ses değişimleri, heceler gibi soyut birimlerle ilgilenir. Sesler konuşma dilinin ayrılabilen parçalarıdır. Diller kendine özgü sınırlı ses dağarcığına sahipken, insan konuştuğu dilin sahip olduğu seslerden daha fazlasını çıkarabilme yeteneğine sahiptir. Bu bakımdan çıkarılabilecek sesler (fon, allofon) sonsuz sayıda olabilir ancak anlam ayırıcı seslerin (sesbirim, fonem) sayısı sınırlıdır. Örneğin <gül> ve <kül> sözcüklerindeki /g/ ve /k/ sesleri iki sözcüğü anlam olarak ayırdığı

için bu sesler fonem olarak adlandırılır. Sesbirimler anlam ayırıcılık özelliğinden ötürü belli bir dile özgü iken seslerin böyle bir özelliği yoktur.

Standart dilde bazı sesler sesbirim olarak tanımlanmadığı halde, bu sesler ağızlarda anlam ayırıcı olarak kullanılabilir. Örneğin açık e denilen /*e*/ sesiyle söylenen "el (organ)" sözcüğü ile kapalı e denilen /*ê*/ sesiyle söylenen "el (yabancı)" sözcüğü ağızlarda farklı söylenirken standart dilde eşsesli kabul edilir [4]. Aynı şekilde /*n*/ ve /*n̂*/ sesleri Türkçede eşsesli kabul edilirken, ağızlarda normal ve genizden olmak üzere birbirinden ayrılan sesbirimleridir [5].

Yazıyla gösterilebilen seslere parçalı sesbirimler (segmental, fonem) denilmektedir. Ciğerlerden gelen havanın ses yolunda bir engelle karşılaşıp karşılaşmaması, gırtlakta meydana gelen daralma, kapanma veya sızma olayları, dudakların şekli, dilin pozisyonu gibi durumlar seslerin kimliğini belirlemektedir. Ses yolunda bir engelle karşılaşmadan ses tellerinin titreşmesi sonucu oluşan seslere ünlü denir. Ünlüler, ses çıkarılırken oluşan ağız boşluğuna, dilin ve dudakların durumuna göre birbirinden ayrılırlar. Ses yolunda tıkanma, daralma, sızma, sürtünme gibi engeller sonucunda oluşan seslere ise ünsüz denir. Ünsüzlerin söylenişinde en az iki organ aktiftir ve bu organların pozisyonlarına göre ünsüzler birbirinden ayrılırlar. Türkçenin ağızları ünlü-ünsüz değişimi bağlamında birbirinden ayrılmaktadır.

Yazıyla (simge) gösterilemeyen ancak söyleyişte var olan seslere parçalar-üstü sesbirimler (supra-segmental, prozodi, bürün) denir. Ton, vurgu, ezgi (melodi), uzunluk (süre) gibi öğeler parçalar-üstü sesbirimlerdir.

Parçalar-üstü sesbirimler yani prozodi konuşma dillerinin vazgeçilmez ve önemli ögelerindendir. Her dilin olduğu gibi ağızların da kendine özgü prozodisi vardır. Bu bakımdan farklı bölgelerde konuşulan Türkçenin ağızları prozodi açısından birbirinden ayrılabilmektedir.

Parçalı sesbirimler alfabede harflerle gösterilir. Standart Türkçede her sesbirimine karşılık bir harf yoktur. Türk alfabesinde 29 harf bulunmasına karşın 35 sesbirim olduğu, ödünç ve parçalar-üstü sesbirimler de eklenince bu sayının 45'e çıktığı gösterilmiştir [6]. Ancak Türkçedeki sesbirimlerin sayısı konusunda tam bir anlaşma yoktur.

Doğal dillerdeki sesler her dil için o dile özgü kurallar çerçevesinde art arda dizilerek önce hece ve daha sonra anlamlı bir yapı oluştururlar. Bir nefeste çıkarılan ses veya seslere hece denir. Türkçenin hece çekirdeğinde (syllable nuclei) ünlüler vardır. Hece yapısı hem fonotaktik hem de prozodik açıdan önemlidir.

Türkçenin ağızlarında standart Türkçede olmayan ünlü ve ünsüz sesbirimler bulunmaktadır. Bu sesbirimler ağızdan ağza da değişmektedir. Bu değişimler 3. bölümde konu edilecektir.

3. TÜRKÇENİN AĞIZLARI

Bu bölümde Türkçenin ağızlarındaki değişimlere ilişkin bilgiler verilecektir. Ağız tanımı yapıldıktan sonra ağızların standart dilden ve birbirilerinden farkları üzerinde durulacaktır.

Ağızlar, dilin yerel konuşma biçimleri olarak tanımlanabilir [7]. Aynı yörede doğup büyüyen insanların konuşma biçimleri birbirine benzer. İnsanın dil edinme süreci kendi ailesinin konuştuğu ağızla başlar, daha sonra eğitimle ve televizyon-radyo gibi araçlarla standart dilin edinilmesiyle devam eder. Standart dil, ülkede konuşulan bir ağzın resmi dil olarak benimsenmesine denilmektedir. Türkiye için İstanbul ağzı standart Türkçe olarak belirlenmiştir. İletişim ve eğitimin kolaylığı açısından bir ağzın standart dil olarak belirlenmesi, diğer ağızların yanlış veya bozulmuş olduğunu göstermez [8].

Türkiye'nin ağızlarıyla ilgili envanter çıkarma çalışmaları 1933 yılına kadar gitmektedir. Yazı dilinde olmayan sözcüklerin derlenmesi amacıyla 150 binden fazla sözcük içeren Türkiye'de Halk Ağızları Söz Derleme Dergisi yayımlanmıştır. Bu derginin bazı eksikleri nedeniyle 1952 yılında başlayıp 8 yıl süren ve 450 bin sözcük içeren Derleme Sözlüğü oluşturulmuştur. Korkmaz'ın [9], ağızlarla ilgili bilgi vermenin yanı sıra ilk fonetik çalışmayı yaptığı belirtilmektedir [10]. 1970'ten sonra Anadolu ağızlarının şehir ve daha küçük birimler özelinde incelendiği çalışmalar yoğunlaşmıştır. Bu çalışmalar daha önce, ilgili ağza özgü sözcüklerin tespit edilmesine yönelik iken daha sonra ses değişimlerini de kapsamıştır. Bu çalışmaların nihai sonucu ağızların sınıflandırılmasıdır. Bu konudaki en son çalışma 1996 yılında Karahan [11] tarafından yapılmıştır. Çalışmada Anadolu ağızları Batı, Doğu ve Kuzey Doğu ağızları olmak üzere 3 ana gruba ayrılarak her bir grubun ayırt edici özellikleri üzerinde durulmuştur [10].

Ağızlar hem kendi arasında hem de standart dilden farklılık gösterir. Bu farklılık kimi zaman sesler (fonetik) bakımından, kimi zamansa ses dizimi, tonlama ve vurgu özellikleri (fonolojik) bakımından kendini gösterir. Bunların yanında şekil (morfolojik), sözcük (leksikal) ve söz dizimi (sentaks) farklılıkları da görülmektedir. Ağızlar konuşmaya dayalı biçimler olduğundan, insanlar bir sesli ifadeyi ağızdan en rahat şekilde çıkartmak eğilimindedirler.

Ses değişmelerine ünlü-ünsüz değişimleri örnek verilebilir. Örneğin /*ñ*/ ünsüzünün /*n, g, ğ, v, y*/ gibi seslere dönüştüğü bilinmektedir. Doğu Anadolu ve Doğu Karadeniz Bölgesi ağızlarında /*ñ*/ ünsüzü /*n*/ olarak değişirken (bildün, gızin, yapdun), Orta Anadolu'da /*ğ*/ şeklini (bağa, öğüne) almaktadır. Bazı ağızlarda ise /*ñ*/ ünsüzü /*y*/ şeklini almaktadır (eliyizden < eli*ñ*izden, durduyuz < durdu*ñ*uz). Bu bakımdan, sadece /*ñ*/ ünsüzündeki değişimin bile, Türkçenin ağızlarını sınıflandırmada kullanılabileceği belirtilmiştir [12].

Aynı şekilde ünlü değişimi için örnek vermek gerekirse; standart Türkçedeki "kazak" sözcüğü, bir ağızda "gazak", başka bir yörede "gozak" şeklinde telaffuz edilebilmektedir.

İnsanlar her söyleyişi gerçekleştiremez. Seslerin bir arada bulunmasının şartları vardır. Sesler aynı ya da komşu hecelerde bir araya gelirken birbirilerinden etkilenmektedir. Bu yüzden Türkçenin ağızlarında ses değişimleri, ses düşmesi veya türemesi gibi birçok ses olayı görülmektedir. Aşağıda Demir ve Yılmaz'ın [5] çalışmasından derlenmiş kimi fonetik ve fonolojik değişimler bir arada verilmiştir.

- "g*õ*" ses öbeği Ankara ağızları için tipiktir. *õ* sesi /*ö*/ sesinin bir alofonu olarak seslendirilmektedir. Örneğin g*õ*rdüm, g*õ*zü.
- Türkçenin özelliklerinden birisi ünlü uyumu kuralına sahip olmasıdır. Ünlü uyumu önlük-artlık uyumu ve düzlük-yuvarlaklık uyumu olarak ayrılmaktadır. Her iki uyumun olduğu bir sözcükte, bir hecenin ardından gelen hecede hangi ünlünün olabileceği aşağıdaki çizelgede gösterilmiştir.

Çizelge 3.1 Ünlü uyumu

Önceki hecede	Sonraki hecede	Örnek
a, ı	a, ı	ada, ırak
e, i	e, i	eve, sinek
o, u	a, u	oda, uçak
ö, ü	e, ü	ödev, ütü

Trabzon ağzında her iki uyum da yoktur: gelsan → gelsen, koli → kolu, benum → benim.

Bunun yanında Türkçeden daha ileri uyum gösteren kimi ağızlar da vardır: başdakı → baştaki, yaparkan → yaparken.

- Ünlü daralmasıyla geniş /*a, e, o, ö*/ ünlüleri standart dilde dar ünlüye dönüşürken ağızlarda daha çok eski geniş haliyle kullanıldığı görülmektedir. Örneğin yokarı → yukarı, gözel → güzel değişimleri gözlenirken, ağızlarda daha çok birinci hal geçerlidir.
- Standart dilde artdamak ünlüleri öndamak ünlüsüne dönüşebilir. Örneğin yaşıl → yeşil, karındaş → kardeş, alma → elma gibi. Ancak bu değişimlerin ilk hallerine ağızlarda karşılaşılabilir.
- Aynı şekilde öndamak ünlüleri artdamak ünlüsüne de dönüşebilir. Ağızlarda bu dönüşüm şöyle olabilir: habar → haber, zalım → zalim.
- Yuvarlak bir ünlünün düz ünlüye dönüşmesiyle düzleşme meydana gelir: kapu → kapı gibi. Ağızlarda çok fazla karşılaşılan bir durumdur. Örneğin gavın → kavun, çamır → çamur. Ancak Doğu Karadeniz ağızlarında düzleşme görülmeyebilir: bizum → bizim gibi.
- Başka dillerden alınmış sözcükler Türkçede uzun söylenirken ağızlarda kısa ünlü haline gelir: zalım → z*â*lim, cahal → c*â*hil.
- Türkçeye başka dillerden geçmiş /*r, l*/ ünsüzleriyle başlayan sözcükler, ağızlarda ünlü türemesine uğrar ve sözcüğün başına dar bir ünlü getirilerek söylenir: ilazım → lazım, irezil → rezil.
- Bazı ağızlarda, sayılarda ünsüz ikizleşmesi meydana gelebilir. Örneğin yeddi → yedi, sekkiz → sekiz vb.
- Ağızlarda ünsüz düşmesi gerçekleşebilir: yapıyom → yapıyorum.
- Söz başındaki /*h*/ sesi bazı ağızlarda düşer: Asan → Hasan.
- Türkçede geniz *n*'si, dudak ünlüsü olan geniz *m*'sine dönüşerek dudaksıllaşır. Örneğin ko*ñ*şu → komşu, do*ñ*uz → domuz gibi. Ancak ağızlarda daha çok ilk hali yaygındır.
- İki sesin yer değiştirmesine göçüşme denir. Ağızlarda şu şekilde rastlanır: yarpak → yaprak, kirbit → kibrit. Ayrıca uzak göçüşme de görülebilir: ireli → ileri.
- Ağızlarda /*k*/ sesi ötümlü hale gelebilir: gapı → kapı, gadın → kadın.

- Sözcük vurgularında da ağızlar arasında farklar görülebilmektedir. Örneğin standart Türkçe ile diğer ağızlarda sözcüklerin vurguları farklı hecelerde olabilir:

  Çizelge 3.2 Vurgu yerleri

Standart Türkçe	Ağızlar
soPA	SOpa
ZONguldak	zonGULdak
piDE	Pİde

- Doğu Karadeniz, Kıbrıs, Güneydoğu Anadolu ağızlarında zayıf dildeki bazı özelliklerin baskın dile transfer edilmesiyle ortaya çıkan değişimler de mevcuttur.

Türkçenin ağızları arasındaki fonetik farklılıklar, onların hem akustik hem de fonotaktik (ses dizimi) olarak birbirinden ayrılabileceğini gösterir. Bunun yanında morfolojik yani şekilsel olarak ağızların tanımından kaynaklanan farklar da vardır. Bu morfolojik farklar da yine fonotaktik açıdan ortaya çıkarılabilmektedir.

Türkçenin ağızlarının birbirinden hız, vurgu, tonlama gibi parçalar üstü özellikleri bakımından ayrıldıkları bilinmektedir. O halde prozodik özelliklerle ağız farklılıkları ortaya konulabilir. Bu çalışmada akustik, fonotaktik ve prozodik bilgi türleri kullanılarak Türkçenin ağızlarının otomatik olarak birbirinden ayrılması üzerinde durulmuştur.

4. DERİN ÖĞRENME

Derin öğrenmenin temelini sinir ağları oluşturur. İnsanın sinir hücrelerinden ilhamla [13] başlayan ve 1950'lerden itibaren algılayıcı [14] olarak anılan sinir ağları zamanla bugünkü anlamını kazanmıştır. Sinir ağları bu süreçte, birkaç defa popüler hale geldiği dönemlerden geçmiştir. Ancak her defasında, belli zorlukları aşamadığı görülmüş ve beklentileri karşılamamıştır. Özellikle 1986'da Rumelhart ve ark. [15] tarafından Geri yayılım (Back-propagation, BP) algoritmasının sinir ağlarında kullanılması gösterilerek gizli katmanlar içindeki etkileşimler deneylerle ortaya konulmuştur. Ancak gizli katmanların eklenmesiyle karmaşıklaşan hesaplamalar ve ihtiyaç duyulan büyük çaplı veriler, zamanın zorluklarını teşkil etmektedir. Sinir ağları 2006'daki Hinton ve ark. [16] çalışmasından sonra popülerliğini tekrar geri almıştır. Sırasıyla Bengio ve ark. [17], Bengio ve LeCun [18], Cireşan ve ark. [19], Krizhevsky ve ark [20], Pascanu ve ark. [21] çalışmalarıyla da bugünkü anlamda "derin" sıfatını kazanmıştır.

Sinir ağı modelinin katman sayısının artması; veri kümelerinin ve bu veriyi işleyebilecek bilgisayar gücü ve belleğinin artışı sonucunda olmuştur. Özellikle, sinir ağlarının dağıtık yapısına daha çok uyan grafik işlemcilerin (GPU) yaygınlaşması derin mimarilerin gerçeklemesini mümkün kılmıştır.

Derin sinir ağlarının yeniden popüler olmasında geliştirilen tekniklerin de katkısı yadsınamaz. Bu yeni tekniklere; ağ ağırlıklarını ilklendirme yöntemleri [22], genelleştirme hatalarını düşürücü yöntemler (regülerizasyon) [23], eniyileme algoritmaları [24], aktivasyon fonksiyonları [25] gibi sinir ağının daha iyi öğrenmesini sağlayan teknikler örnek verilebilir. Derin sinir ağlarının, sınıflandırma işlemi ile özniteliklerin çıkartılması işlemini birleştirdiği düşünülebilir [26]. Her katmanda probleme ait öznitelikler öğrenilir ve öğrenilen bu öznitelikler bir sonraki katman için girdi oluşturur. Böylelikle giriş katmanından çıkış katmanına doğru en basitten en karmaşık özniteliklerin öğrenildiği bir yapı kurulmaktadır [27].

O halde derin öğrenme, derin mimariye dayalı sinir ağları ve bunlar için geliştirilen yöntemler bütünü olarak kısaca tarif edilebilir [28].

Sinir ağlarının derinleşerek popüler hale gelmesiyle birlikte, 1990'larda icat edilen sinir ağı mimarileri de daha çok kullanılmaya başlamıştır. Bu mimarilerden ikisi dikkat çekicidir. Bunlardan biri, 1991'de Hochreiter [29] tarafından geliştirilen yinelemeli sinir ağlarıdır (Recurrent Neural Network, RNN). Bu ağlar dizi modelleme amacıyla geliştirilmiş ve uzun dönemli veriyi modellemede kullanılmıştır. Ancak bazı nedenlerle bu ağların öğrenme kabiliyetinin pratikte sınırlı olduğu görülmüştür. Bu zorluğun üstesinden 1997'de Hochreiter ve Schmidhuber [30] tarafından geliştirilen uzun kısa-dönem bellekli (Long Short-Term Memory, LSTM) yinelemeli sinir ağlarıyla gelinmiştir. Bu mimari günümüzde konuşma tanıma ve doğal dil işleme gibi zamansal dizi modelleme problemlerinde sıklıkla kullanılmaktadır.

Önemli mimarilerden bir diğeri konvolüsyonel sinir ağlarıdır (Convolutional Neural Networks, CNN). LeCun ve ark. [31] tarafından günümüzdeki görünümüne kavuşan bu mimari, iki boyutlu veri (görüntü, ses, vb.) üzerinde başarıyla uygulanmıştır.

Özel yapılı derin sinir ağlarına geçmeden önce bu tez çalışmasında da kullanılan geleneksel ileri beslemeli sinir ağlarını tanıtmak ve temel oluşturan hesaplamaları vermek faydalı olacaktır.

4.1 İleri Beslemeli Sinir Ağları

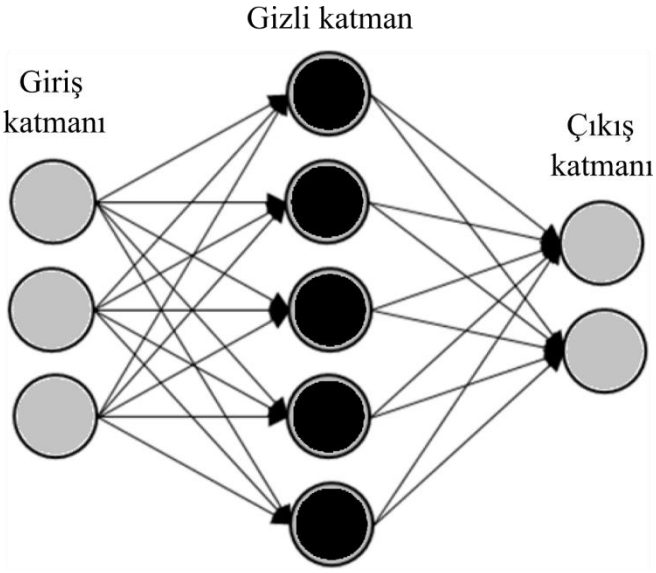
Sinir ağları verilen bir girdi verisini (x_i), bir çıktıya (y_i) haritalamaya yarayan işlem birimleridir. Burada (x_i, y_i) çifti herhangi bir probleme özgü olabilir. Örneğin x bir görüntünün pikselleri ve y de görüntüde tespit edilmesi gereken bir nesne olabilir. Ya da x bir konuşma sinyali iken, y konuşmayı oluşturan fonemler olabilir.

Gözetimli öğrenme şeklinde çalışan ileri beslemeli sinir ağları (x, y) eğitim verisini kullanarak x'i y'ye ilişkilendiren bir model geliştirir. x ile y arasında doğrusal veya doğrusal olmayan bir ilişki olabilir.

İleri beslemeli sinir ağlarına çok katmanlı algılayıcılar (Multi Layer Perceptron, MLP) da denilmektedir. Ancak algılayıcıların aktivasyon fonksiyonları doğrusal olduklarından doğrusal olarak ayrılamayan veri kümelerinde başarısız olmaktadır. Doğrusal aktivasyon fonksiyonu [0,1] gibi ayrık değerli, belli bir eşik değerin üstündekileri 1, diğerlerini 0

olarak çıkışa veren bir fonksiyondur. İleri beslemeli sinir ağları ise bunlardan farklı olarak aktivasyon için genelde sigmoid fonksiyonları kullanır. Sigmoid fonksiyonları doğrusal değildir ve sürekli (continuous) değerler üretir.

İleri beslemeli sinir ağları giriş ve çıkış katmanlarından başka, bir veya birden fazla gizli katmana sahiptir. Bu katmanlarda bulunan temel işlem birimlerine nöron denir. Bütün katmanların bağlantıları tek yönlüdür ve çıkış katmanına doğrudur. Tek yönlü özelliğinden ötürü bu tip sinir ağlarında bir döngü oluşmaz. Bu şekilde bir katmanın çıktısı bir sonraki katmanın girdisi olmaktadır. Bir katmandaki bütün nöronların bir sonraki katmanda bulunan bütün nöronlara bağlandığı bu tipteki bağlantı için tam bağlı (fully connected, dense) katman ifadesi kullanılır. Tek gizli katmana sahip bir ileri beslemeli sinir ağı şekilde gösterilmiştir.

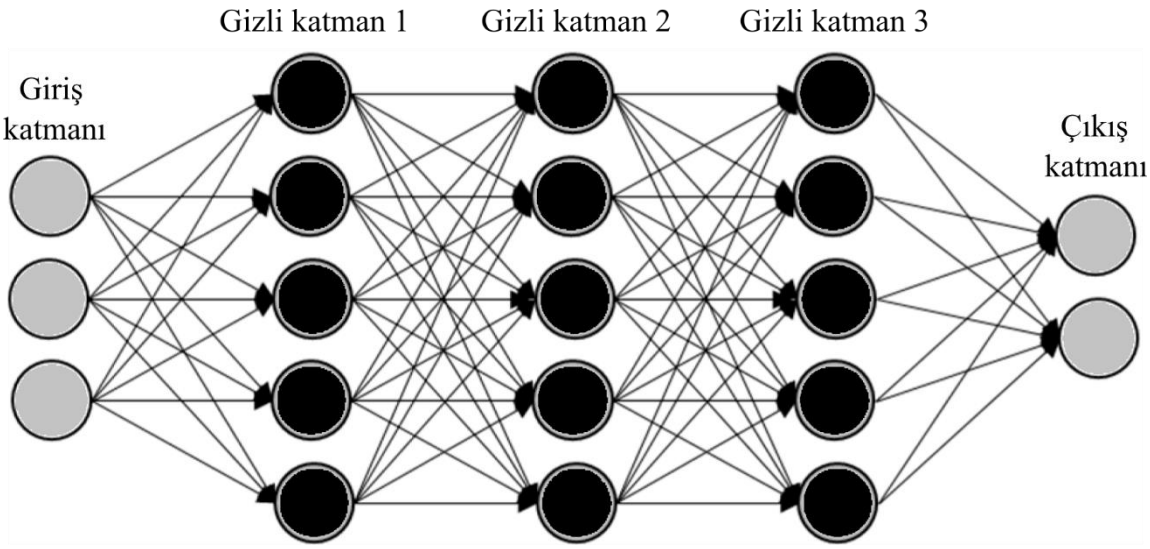


Şekil 4.1 Tek gizli katmanlı ileri beslemeli sinir ağı.

Evrensel yakınsama teoremine [32] göre bir gizli katmana sahip sinir ağı herhangi bir fonksiyonu yakınsayabilir. Bu bakımdan çok katmanlı ileri beslemeli sinir ağlarının istenilen bir fonksiyonu yakınsayabildiği düşünülmektedir.

Genelde ikiden fazla gizli katmanı olan sinir ağları için “derin” ifadesi kullanılmaktadır. Ancak bir sinir ağının “derin” vasfını kazanması için gerekli gizli katman sayısında bir

uzlaşma olmadığını da belirtmek gerekir. Aşağıdaki şekilde derin, çok katmanlı sinir ağı görülmektedir. Derin sinir ağları verideki karmaşık ilişkileri modelleyebilir.



Şekil 4.2 Üç gizli katmanı bulunan beş katmanlı derin sinir ağı.

Öğrenme, veri işlendikçe sinir ağının çıkışındaki hatayı azaltacak şekilde, bağlantı ağırlıklarının değiştirilmesi anlamına gelir. Başlangıçta sinir ağının nöronları arasındaki bağlantılara rastgele değerler (ağırlık) verilir. Girdi ile ağırlıklar çarpılıp toplanarak bir fonksiyondan geçirilir ve sürekli değere sahip bir çıktı elde edilir. Bu işlem çıkış katmanında da tekrar edilerek girdi için sonsal olasılık değeri hesaplanır. Ağın ürettiği bu olasılık değeri gerçek çıktı ile karşılaştırılarak aralarındaki hata miktarı bulunur. Eğer bu hata miktarı yüksekse, sinir ağının kendisinden beklenen davranışı sergilemediği düşünülür ve hatayı düşürecek adımlar atılır. Bunun için nöronlar arasındaki ağırlık değerleri değiştirilir. Ağırlık değerlerinin ne ölçüde değiştirileceğini hesaplamak için ise hatanın her bir ağırlık değerine göre türevi alınır. Böylece ilgili ağırlık değerinin hatadaki rolü ölçülerek ağırlık değeri buna göre biraz arttırılır veya azaltılır. Bu güncel ağırlık değerleriyle ağın eğitimi başa döner ve işlemler, hata en düşük seviyeye gelene kadar tekrar eder. Hata en düşük seviyeye geldiğinde sinir ağı, x ile y arasındaki ilişkiyi öğrenmiş ve başarılı bir haritalama yapıyor demektir. Burada doğru haritalamayı sağlayan yegane faktörün nöronlar arasındaki doğru ağırlık değerleri olduğu söylenebilir. Bu da eğitim süresince elde edilmektedir. Burada sözel olarak ifade edilen, sinir ağının eğitimi sürecindeki matematiksel işlemler sırasıyla aşağıda verilmiştir.

Sinir ağının giriş katmanı verinin doğrudan bağlandığı ve herhangi bir işlem yapılmayan katmandır. Bu yüzden matematiksel işlemleri giriş katmanı dışındaki nöronlar gerçekleştirir. Bir nöronda matris vektör çarpımı yapıldıktan sonra elde edilen sonuca bir aktivasyon fonksiyonu uygulanır. Örneğin tek gizli katmanlı bir sinir ağında (Şekil 4.3), gizli katmanın girişine bağlı olan ağırlıklar ile girdi vektörü çarpılır:

$$z^{(1)} = W^{(1)}x + b \tag{4.1}$$

Burada $W^{(1)}$ giriş ile gizli katman arasındaki ağırlık matrisini, x girişe verilen öznitelik vektörünü, b yanlılık (bias) vektörünü, $z^{(1)}$ ise toplam vektörünü gösteren parametrelerdir. Matris vektör çarpımıyla elde edilen z vektörüne g aktivasyon fonksiyonu uygulanarak sinir ağının doğrusal olmayan bir nitelik kazanması sağlanır.

$$a = g(z^{(1)}) \tag{4.2}$$

Aktivasyon fonksiyonu olarak genelde hiperbolik tanjant veya lojistik gibi sigmoid karakterli fonksiyonlar veya son yıllarda sıkça rastlanan ReLU [25] fonksiyonu kullanılmaktadır. g fonksiyonu z'nin her bir değerine ayrı ayrı uygulanır. Bu aşamada a, fonksiyondan elde edilen, sürekli değerler içeren bir vektördür ve gizli katmanın çıkışıdır. Bundan sonra sinir ağının çıkış katmanına bir girdi oluşturarak bu sefer gizli katman ile çıkış arasındaki ağırlık matrisi ($W^{(2)}$) ile çarpılır. Elde edilen $z^{(2)}$ toplam vektörüne bu sefer f softmax fonksiyonu uygulanır.

$$z^{(2)} = W^{(2)}a + b \tag{4.3}$$

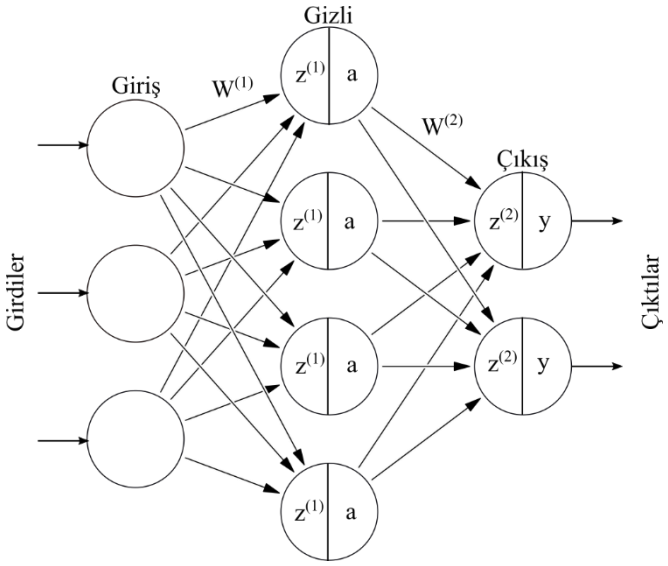
$$y = f\left(z^{(2)}\right) \tag{4.4}$$

Burada y sinir ağının çıktısını oluşturmaktadır. Sinir ağı, ayırıcı (discriminative) yolla eğitildiği için softmax fonksiyonu sonsal (posterior) olasılıkların hesaplanmasını sağlar. Sonsal olasılık her bir girdi için sinir ağının tahminlerini ifade eder ve $P(y|x)$ koşullu olasılığı ile gösterilir.

4.1.1. Sigmoid Fonksiyonları

Algılayıcı ağlarda kullanılan adım (step) fonksiyonu doğrusal olarak ayrılamayan veri kümelerini sınıflandırmada başarısızdır. Bu nedenle doğrusal olmayan problemler için sinir ağlarında sigmoid fonksiyonları kullanılmaktadır. Lojistik ve tanjant fonksiyonlarına S şekillerinden ötürü sigmoid karakterli fonksiyonlar denir. Lojistik fonksiyonu kendisine

girdi olarak verilen vektörün her bir değerine ayrı ayrı uygulanır ve onları 0 ile 1 arasında bir değere çekerken, tanjant fonksiyonu girdiyi -1 ile $+1$ arasına haritalar.



Şekil 4.3 İleri beslemeli sinir ağının gizli katman ve çıkış işlemleri.

$$g_{log}(z) = \frac{1}{1 + e^{-z}} \tag{4.5}$$

$$g_{tanh}(z) = \frac{e^{z} - e^{-z}}{e^{z} + e^{-z}} \tag{4.6}$$

Çıkış katmanında lojistik fonksiyonu kullanılabilir. Özellikle bir girdi için birden fazla etiket (multi label) içeren problemlerde (örneğin bir görüntüdeki nesnelerin tespitinde veya bu tez çalışmasının 5.4 Kısmında anlatılan dil modellemesinde) lojistik fonksiyonu etkili olmaktadır. Genelde çok sınıflı ancak tek etiketli veri kümelerinde lojistik fonksiyonunun genelleştirilmiş hali olan softmax fonksiyonu kullanılmaktadır [33].

4.1.2. Softmax Fonksiyonu

Sinir ağının çıkışında, girdiye karşılık gelen sonsal olasılıkları elde etmek için softmax fonksiyonu kullanılır. Girdi olarak verilen veriyi 0 ile 1 arasında, toplamı 1 olan olasılıklara haritalar. Böylece her bir girdi için sinir ağının tahminini oluşturur.

$$P(y = k|x) = \frac{e^{f_k(x)}}{\sum_{j=1}^{K} e^{f_j(x)}} \tag{4.7}$$

Burada k adet sınıf için $f_k(x)$ sonsal olasılıkları gösterir, dolayısıyla toplam içindeki payını ifade eder. Sinir ağının çıkış katmanında k adet sınıfı gösteren sayıda nöron

bulunur. Softmax fonksiyonu eğitim verisindeki gerçek etiketlere karşılık gelen nöronların çıktısını 1'e yakınlaştırarak olasılığı maksimize etmeye yarar. Dolayısıyla çok sınıflı tek etiketli problemlerde sonsal olasılığı elde etmek için kullanılması yaygındır. Ancak çıkış katmanında sigmoid fonksiyonu kullanıldığında softmax gibi tek bir çıktıyı değil, birden fazla çıktının değerini birbirinden bağımsız olarak maksimize edebilmektedir.

4.1.3. Maliyet Fonksiyonu

Sinir ağının tahmini elde edildikten sonra, eğitim verisinin gerçek etiketleriyle karşılaştırılarak hata miktarının belirlenmesi gerekir. Bunun için genelde kare hata (squared error) veya çapraz entropi (cross entropy) maliyet fonksiyonları kullanılır. Bu tez çalışmasında çapraz entropi maliyet fonksiyonuna göre sinir ağının performansı ölçülmüştür. Softmax fonksiyonu ile elde edilen p olasılıkları ile N adet eğitim verisindeki gerçek y değerleri arasındaki hata miktarı çapraz entropi ile şöyle hesaplanır:

$$C = -\frac{1}{N}\sum_{i=1}^{N} y_i \log_2 p_i \tag{4.8}$$

Sinir ağı, x girdisine karşılık gelen y çıktısı arasında haritalama yapmaya çalışır. İlk iterasyonda, ağın başlangıç parametreleri rastgele dağıtıldığı için hata miktarı yüksek çıkar. Bu yüzden hatanın düşürülmesi gerekir. Bunun için mevcut hata miktarına neden olan ağırlık değerleri değiştirilmelidir. Mevcut hatanın ağırlık değerlerine göre türevleri alınırsa, ağırlık değerlerinin mevcut hatadaki katkısı hesaplanabilir. Buna göre ilgili ağırlık değeri çok küçük miktarda arttırılıp azaltılarak güncellenir.

Birden fazla etiketi olan örnekler için çıkış katmanında sigmoid fonksiyonu ile birlikte ikili çapraz entropi (binary cross entropy) maliyet fonksiyonu kullanılmaktadır. Bu tez çalışmasının 5.4 Kısmında bu konuya değinilmiştir.

4.1.4. Geri Yayılım Algoritması

Türev alma işlemi için geri yayılım (Back Propagation, BP) algoritması kullanılır. Dolayısıyla geri yayılım algoritmasının amacı değişimin yönünü ve miktarını belirlemektir denilebilir. Birden fazla gizli katman olduğu durumlarda zincir türev alma işlemi gerçekleştirilir. ∇F toplam hatanın ağırlık değerlerine göre türevini ifade eder.

$$\nabla F\left(w_{ij}^{n}\right) = \frac{\partial C}{\partial w_{ij}^{n}} \tag{4.9}$$

Bu şekilde değişimin yönü ve miktarı belirlendikten sonra ağ parametreleri buna göre güncellenebilir. Güncel parametrelerle ağın eğitimi tekrar başlar ve hata olabilecek en düşük düzeye gelene kadar devam eder. Bu sürece eniyileme (optimization) denilmektedir. Eniyileme için genelde gradyan temelli algoritmalar (SGD [34], Adam [24]) kullanılır.

4.1.5. Eniyileme

Sinir ağının en iyi parametreleri edinmesi için eniyileme algoritması ile eğitilir. Eğitim süresince ağ parametreleri güncellenir. Güncelleme işlemi her örnekten sonra olabileceği gibi seçilen bir mini-yığın (mini-batch) kadar örnek işlendikten sonra da olabilir. Güncelleme işlemi aşağıdaki gibi olur:

$$w_{ij}^{n+1} = w_{ij}^{n} - \propto \nabla F\left(w_{ij}^{n}\right) \tag{4.10}$$

Burada $n + 1$ bir sonraki iterasyonu, $\propto$ öğrenme oranı (learning rate) katsayısını gösterir. Öğrenme oranı ile türevlerin kontrol edilmesi sağlanarak güncelleme işleminin istenilen miktarda olması garanti edilmektedir. Öğrenme oranı çok düşük seçildiğinde sinir ağının yakınsaması gecikmekte, büyük seçildiğinde ise güncelleme değeri sürekli salınarak optimizasyon kontrolden çıkmaktadır. Bunun üstesinden gelmek için $\propto = 0{,}1$ gibi bir değer seçilir ve doğrulama (validation) verisi üzerinde ağın performansı test edilir. Belli bir örnek sayısından sonra doğrulama verisinin olasılık değerlerinin artması durumunda eğitim devam eder, aksi halde $\propto$ değeri belli oranda düşürülür. Olasılık değerinin önemli oranda artmaması halinde ise eğitim sonlandırılır.

4.1.6. Nöron Düşürme

Son yıllarda derin öğrenmenin popüler hale gelmesiyle birlikte yeni teknikler de geliştirilmiştir. Bunlardan biri nöron düşürme (dropout) [23] denilen regülerizasyon yöntemidir. Bilindiği gibi sinir ağlarında aşırı öğrenme (overfitting) sorunu görülmektedir. Sinir ağı eğitim örneklerini bir anlamda ezberlemekte ve önceden görmediği örnekler üzerinde iyi tahmin yapamamaktadır. Aşırı öğrenme sorununun üstesinden gelmek için geliştirilen nöron düşürme yönteminde, her güncellemede katmanlardaki nöronlar belli bir yüzde miktarıyla rastgele seçilerek diğer katmanlarla olan bağlantıları koparılır. Bu sayede,

kalan diğer nöronların da sonuca etki etmesi sağlanır. Böyle yapılarak sadece bazı nöronların değil, bütün nöronların iyi öğrenerek tahminde etkili olması hedeflenmektedir.

4.1.7. Ağırlık İlklendirme

Sinir ağının derinliği arttıkça ağırlık matrislerine uygun başlangıç değerleri atanması gereği görülmüştür [22]. Böyle yapılarak hem hızlı bir şekilde yakınsama yapılması hem de hata türevinin azalarak yok olması (vanishing gradient) veya aşırı artması (exploding gradient) gibi sorunların önlenmesi sağlanır [35].

Ağırlık ilklendirme için değerler, örneğin ortalaması sıfır ve varyansı $\sigma^2 = \frac{2}{n_{in}+n_{out}}$ olan bir dağılımdan alınmaktadır [22]. Burada n, ilgili katmanı besleyen ve o katmanın beslediği nöron sayılarını ifade eder.

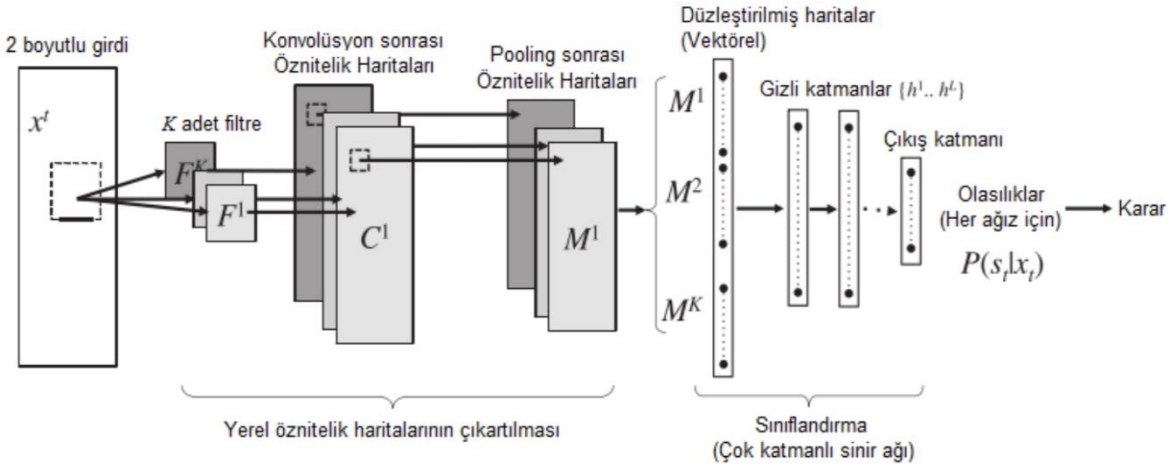
4.2. Konvolüsyonel Sinir Ağları

Konvolüsyonel Sinir Ağları [31], görüntü tanıma [36], akustik sinyal işleme [37, 38], dil tanıma [39, 40], duygu tanıma [41], konuşmacı doğrulama [42] uygulamalarında sıklıkla kullanılmaktadır. Bu ağlar iki aşamalı çalışır. Birinci aşamada birbiriyle ilişkili olan yerel özniteliklerin çıkartılmasını, ikinci aşamada ise çok katmanlı sinir ağları kullanılarak sınıflandırma yapılmasını sağlar. Bu aşamalar 2-boyutlu girdi üzerinde (Şekil 4.4) gösterilmiştir.

4.2.1. Konvolüsyon

Küçük boyutlu filtreler (kernel) girdinin tümü üzerinde gezdirilir. Bu gezinme (konvolüsyon) sırasında, girdi verisinin ilgili pozisyonundaki değerler ile filtre içindeki değerler eleman elemana çarpılarak toplanır (Şekil 4.5). Bu işlem girdinin başından sonuna kadar uygulandığından belli bir özniteliğin girdide var olup olmadığını, varsa girdinin neresinde olduğunu belirlemeye yarar.

Bu yüzden, bu işlemin sonucunda girdinin bir filtreden geçirilmiş yeni bir görünümü (temsili) ortaya çıkar. Bu yeni görünüme öznitelik haritası (feature map) adı verilir (Şekil 4.5).



Şekil 4.4 İki boyutlu girdi verisiyle basit CNN mimarisi [43].

Öznitelik haritası girdinin içindeki komşuluk ilişkilerini korumaktadır. Bir girdiye konvolüsyon katmanında ne kadar filtre uygulanırsa, o kadar sayıda öznitelik haritası çıkar, dolayısıyla da o kadar sayıda öznitelik elde edilir. Ancak filtre sayısı arttıkça işlem maliyeti de artacağından mümkün olan en az sayıda filtre uygulanması istenir [44].

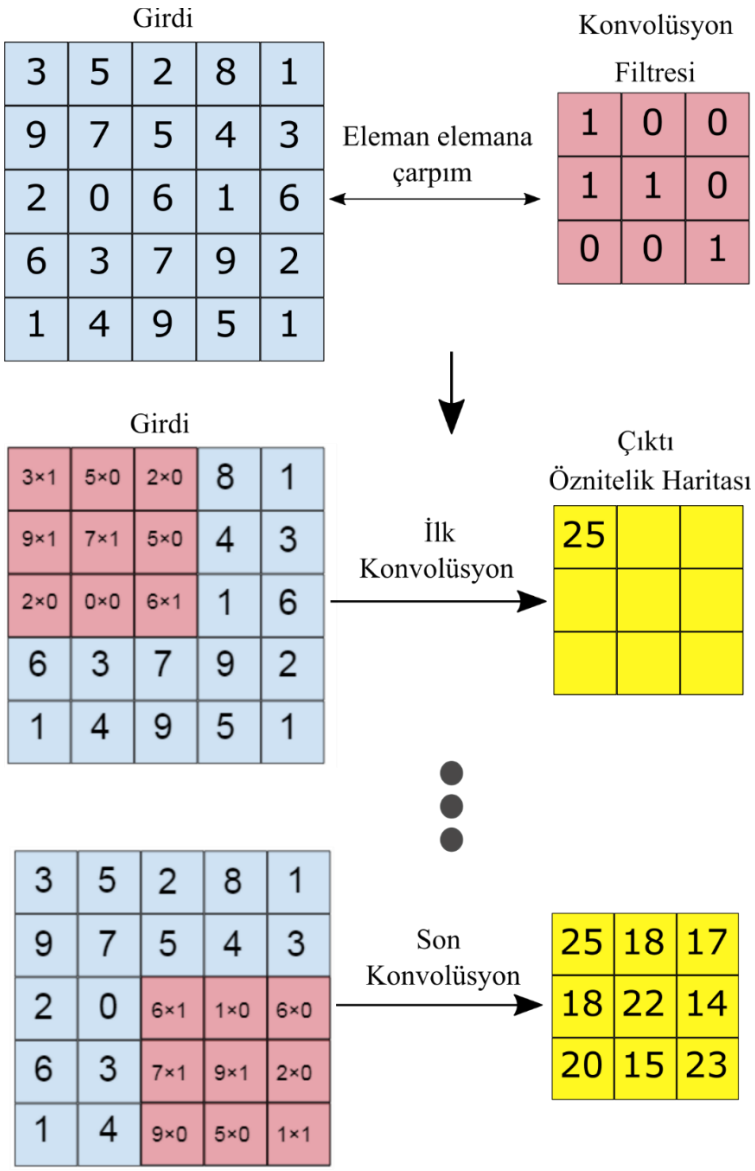
Bir filtrenin içindeki değerler sinir ağının bağlantı ağırlık (weight) değerlerine karşılık gelir. Ayrıca girdi verisinin filtre ile çarpılan kısmına algı alanı (receptive field) adı verilir. Bir algı alanı konvolüsyon katmanındaki bir nörona bağlıdır. Dolayısıyla konvolüsyon katmanındaki bir öznitelik haritasına ait bütün nöronların girdi verisiyle arasındaki bağlantı ağırlıkları aynıdır (weight sharing). Bundan dolayı girdinin farklı yerlerinde aynı öznitelik aranmış olmaktadır.

Başlangıçta filtrelerin parametreleri (ağırlık matrisi) rastgeledir ancak daha sonra eğitim boyunca filtre içeriği güncellenerek ideal değerlerine ulaşır.

$G \times G$ ebatlı bir girdi üzerinde $D \times D$ boyunda bir filtre, adım boyu s (Şekil 4.5, $s = 1$, matris üzerinde sağa ve aşağıya doğru atılacak adım boyu.) olacak şekilde kaydırıldığında oluşan öznitelik haritasının boyu $(G - D)/s + 1$ hesabıyla bulunur. Eğer filtre kare şeklinde değilse, ilgili eksen üzerinde işlem yapılır. Örneğin spektrogram üzerinde tek bir yönde, yani frekans veya zaman ekseninden biri boyunca filtreler uygulandığında diğer eksen için kenar uzunluğu (D) 1 olarak alınır.

4.2.2. ReLU Fonksiyonu

Öznitelik haritasının çıkışına düzeltilmiş doğrusal birim (Rectified Lineer Unit, ReLU) [25] aktivasyon fonksiyonu uygulanarak sonuca doğrusal olmayan bir özellik katılmış olur. ReLU fonksiyonu $F(x) = \max(0, x)$, her bir öznitelik değeri üzerinde $x \leq 0$ olan değerler için 0, $x > 0$ olan değerler içinse x değeri döndürmektedir. Bu bakımdan ReLU fonksiyonunun türevini almak oldukça basittir. Sinir ağının eğitimi boyunca hızlı ve kararlı bir şekilde yakınsamayı sağladığı gösterilmiştir [20].

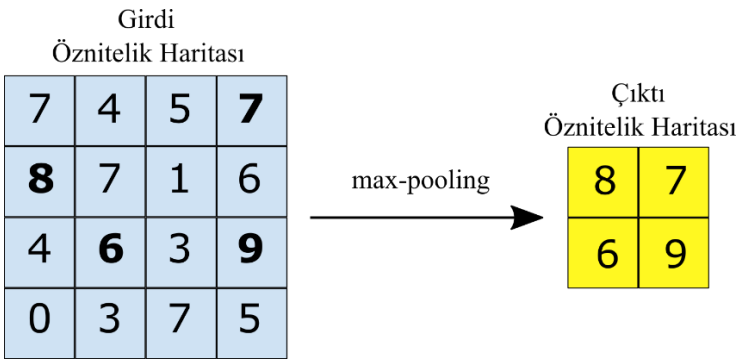


Şekil 4.5 Girdi üzerinde (5×5) konvolüsyon filtresinin (3×3) uygulanmasıyla öznitelik haritasının (3×3) elde edilmesi [45].

4.2.3. Pooling

Elde edilen öznitelik haritaları daha sonra pooling (havuzlama) adı verilen işleme tabi tutulur. CNN mimarisinde önemli bir yer tutan pooling işlemi öznitelik haritalarının boyutunu düşürür ve böylece özniteliklerdeki çeşitlilikleri azaltır. Pooling işlemi için çeşitli kaynaklarda örnekleme ifadesi de kullanılır. Boyut düşmesine rağmen hala önemli öznitelik bilgisi korunmaktadır. Pooling işlemi genelde en büyüğü bulma (max) operatörü ile yapılır. Konvolüsyon filtresinin boyu gibi önceden belirlenen bir pooling penceresinin içerisinde kalan özniteliklerin en büyük elemanı alınır ve öznitelik haritası boyutu düşürülerek tekrar oluşturulur. Şekil 4.6'da 4×4 boyunda bir öznitelik haritasına 2×2 boyunda pooling penceresi uygulanarak öznitelik haritasının boyutu yarıya düşürülmektedir. Pooling penceresinin adım boyu 2'dir ve birbiriyle örtüşmeyecek şekilde öznitelik haritasına uygulanır. Pooling işleminin parametreleri (ağırlık matrisi) olmadığından bu aşamada öğrenme yoktur.

CNN mimarisinin konvolüsyon katmanındaki paylaşımlı ağırlık özelliği ve pooling işlemi; girdideki değişkenliklerin azaltılması, hesaplama gücünden kazanç, özniteliklerde gürbüzlük gibi faydalar sağlar.



Şekil 4.6 Elde edilen öznitelik haritasının max-pooling işlemiyle boyutunun düşürülmesi.

Konvolüsyon ve pooling adımları istenen sayıda art arda eklenerek tekrarlanabilir. Bu durumda bir katmandaki öznitelik haritası bir sonraki katman için girdi oluşturmaktadır. Her konvolüsyon katmanından elde edilen öznitelikler bir önceki özniteliklerden daha karmaşık yapıdadır. Böylece sinir ağının sonuna doğru en genelden en özele olmak üzere öznitelikler öğrenilmektedir.

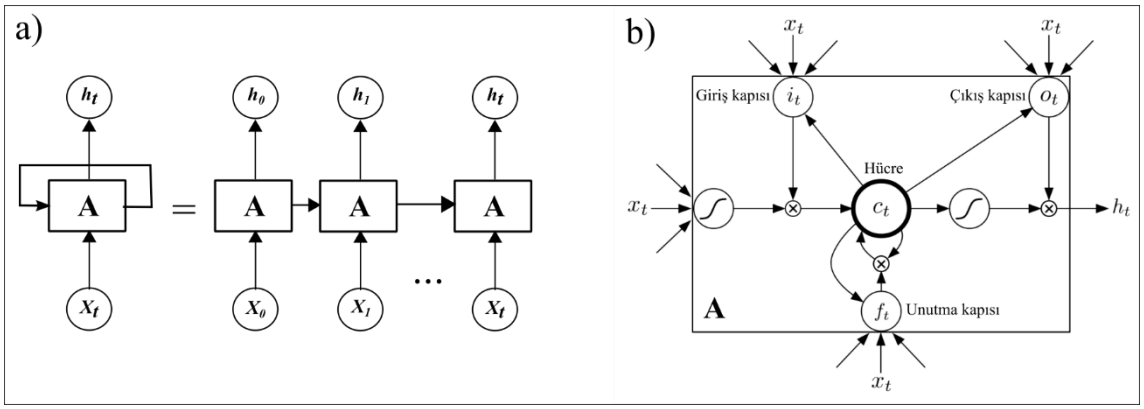
4.2.4. Çok Katmanlı Sınıflandırıcı

CNN mimarisinin sınıflandırma aşaması çok katmanlı (derin) sinir ağlarından oluşur. Burada amaç, elde edilen öznitelikler kullanılarak sınıflandırma işleminin yapılmasıdır. Şekil 4.4'de görüldüğü üzere öznitelik haritaları üç boyutludur. Bu yüzden pooling işleminden elde edilen öznitelik haritaları düzleştirilip vektör haline getirilerek çok katmanlı sinir ağının girişine verilir. Genelde sinir ağının bu kısmındaki gizli katmanlarda sigmoid aktivasyon fonksiyonu, çıkış katmanında ise softmax fonksiyonu kullanılır [38]. Çıkış katmanında her bir sınıf için $P(s_t|x_t)$ sonsal olasılıkları hesaplanır. Bunlardan en yüksek olasılıklı olan seçilerek sınıflandırma tamamlanmış olur.

4.3. Yinelemeli Sinir Ağları

Yinelemeli Sinir Ağları (Recurrent Neural Networks, RNN) ardışık zaman serilerinin modellenmesini sağlayan, klasik sinir ağlarının genelleştirilmiş hali olarak görülebilir. Yinelemeli sinir ağlarının hesaplama birimleri açık hale getirilerek ileri beslemeli sinir ağına dönüştürülebilmektedir (Şekil 4.7a). RNN'ler konuşma ve el yazısı tanıma [46, 47], makine çevrimi [48, 49] ve dil modelleme [50, 51] gibi birçok ardışık yapılı probleme başarıyla uygulanmıştır.

Yinelemeli katmanın çıkış aktivasyonu, bir sonraki zaman adımında katmanın kendi girişine bağlanarak ağın verideki zamansal bağımlılıkları modellemesi sağlanır. Bu, serinin bir zaman adımını işlerken sinir ağının önceki zaman adımlarını hatırlaması demektir. Olasılık modeli açısından bakıldığında, bu yinelemeli bağlantı, modelin mevcut zaman adımındaki tahminini önceki zaman adımlarına göre yapmasını sağlamaktadır. Bunlar tek yönlü (uni-directional) RNN olarak isimlendirilir.



Şekil 4.7 a) Yinelemeli bağlantı ve zaman boyunca açılmış hali. b) LSTM hücresi.

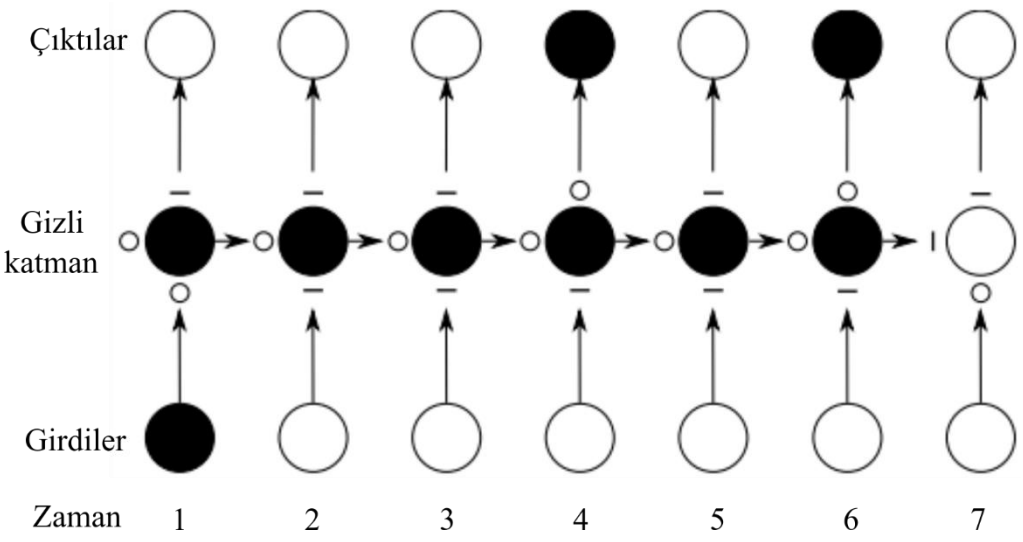
Çift yönlü (bi-directional) RNN'ler, seriyi her iki yönde, başından ve sonundan başlayarak işlemektedir. İleri ve geri yönlü işlemler için birer olmak üzere iki adet tek yönlü RNN'den oluşurlar. İleri ve geri RNN'lerin gizli aktivasyonları uç uca eklenerek birleştirilmektedir. Serinin tamamına erişimin mümkün olduğu yani bütün serinin elde olduğu problemler için uygundur.

Uzun dönemli geçmiş verisinin öğrenilmesi için hatanın ağ içerisinde geriye doğru işletilmesi (back propagation) gerekir. Ancak RNN, türevlerin geriye doğru hesaplanması aşamasında türevin azalarak yok olmasına (vanishing gradient) veya çok yüksek değerlere çıkmasına (exploding gradient) neden olmaktadır [52]. Bu nedenle yinelemeli sinir ağlarının eğitimi pratikte zor hale gelmekte ve 5-10 zaman adımını geçen zamansal bağımlılıkların modellenmesi zorlaşmaktadır [53]. RNN yapısında olan LSTM sinir ağları, RNN'nin bu dezavantajını ortadan kaldırmak üzere tasarlanmıştır. Bunu da ağa eklenen bellek hücreleri ve çeşitli kapılarla sağlar [30]. Böylece bu yapı, zamansal verideki kısa veya uzun örüntüleri öğrenebilir hale gelmektedir.

Şekil 4.7b'de LSTM'in bellekli ve kapılı yapısı görülmektedir. Klasik sinir ağlarındaki gizli katmanların yerini bellek blokları almıştır. Bu bloklar bilgisayarlardaki bellek birimi olarak düşünülebilir. Her blokta yinelemeli olarak bağlanmış bir veya birden fazla bellek hücresi ve üç çarpım birimi (giriş i, çıkış o ve unutma f kapısı) vardır. Bu çarpım birimleri sırasıyla yazma, okuma ve silme işlemlerini gerçekleştirerek bellek hücresinin davranışını kontrol ederler. Giriş kapısı girdi aktivasyonlarının bellek hücresine girişini kontrol ederken; çıkış kapısı, bellek hücresinin çıktı aktivasyonlarının ağın geri kalanına akışını kontrol eder. Unutma kapısı bellek bloğundan bellek hücresine doğru bilgi akışını kontrol ettiğinden hücrenin belleğinin silinmesini (unutmasını) sağlar. Giriş ve çıkış kapıları türevin azalarak yok olmasını engeller. Bununla birlikte unutma kapısı sonsuz döngüyü engelleyerek ağın sınırlı belleğe sahip olmasını sağlar. Şekil 4.7b'de bellek bloğunun çıkışı, bir sonraki adımda hem bloğun girişine hem de kapılara gitmektedir. Kapılar gelen bilginin ne yapılacağına karar vermek üzere işlem (yaz, oku, sil) yaparlar.

Şekil 4.8'de kapıların işlevleri görülmektedir. Gizli katman düğümlerinin üstünde okuma (output), solunda unutma (forget) ve altında yazma (input) kapıları yer almaktadır. Kısa çizgi kapalı, küçük çember ise açık anlamına gelir. Şekil, girişe verilen bilginin gizli

katman boyunca ilerlediğini, 4. ve 6. zaman adımlarında gizli katman durumunun (bellek bloğunun) çıkışa verildiğini, 7 adımda ise bilginin unutma kapısı vasıtasıyla unutulduğunu göstermektedir.



Şekil 4.8 LSTM kapılarının (gates) işlevleri [54].

Buradan hareketle, ileri beslemeli sinir ağının bire-bir haritalama yaparken, yinelemeli sinir ağlarının bire-çok (resimler için açıklama yazısı yazma), çoka-çok (makine çevrimi), çoka-bir (ses tanıma, dil tanıma) haritalama yaptığı söylenebilir.

LSTM katmanının vektör hesaplama işlemleri aşağıda verilmiştir:

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i) \tag{4.11}$$

$$f_t = \sigma\left(W_{xf}x_t + W_{hf}h_{t-1} + b_f\right) \tag{4.12}$$

$$g_t = \tanh\left(W_{xg}x_t + W_{hg}h_{t-1} + b_g\right) \tag{4.13}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t \tag{4.14}$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o) \tag{4.15}$$

$$h_t = o_t \odot \tanh(c_t) \tag{4.16}$$

Yukarıdaki eşitliklerde W ağırlık matrislerini, b bias vektörlerini göstermektedir. x_t ve h_t giriş ve çıkış dizisini; i_t, f_t, o_t sırasıyla giriş, unutma ve çıkış kapılarını temsil etmektedir. g_t girişi ve önceki durumu hesaba katarak şimdiki duruma dönüştürme işlemidir. c_t hücrenin durumunu güncelleme adımıdır. $\sigma(\cdot)$ sigmoid fonksiyonu, Θ elemanlı çarpma

işlemini ifade eder. t ise 1'den T'ye kadar zaman adımlarını göstermektedir. Bu hesaplamalar sonucunda, LSTM ağının çıkışında elde edilen h_t'ye softmax fonksiyonu (ϕ) aşağıdaki gibi uygulandığında sonsal olasılık dağılımı elde edilir:

$$y_t = \phi(h_t) \tag{4.17}$$

1'den T'ye kadar her zaman adımında yukarıdaki işlemler tekrar edilerek $x = (x_1, \dots, x_T)$ giriş dizisinden $y = (y_1, \dots, y_T)$ çıkış dizisine bir haritalama yapılmış olur.

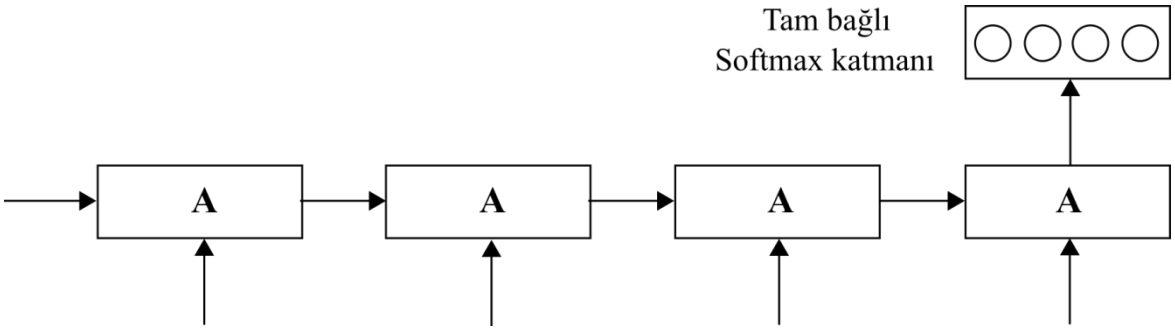
Yinelemeli sinir ağları normal geri yayılım algoritmasıyla eğitilebilir. Ancak böyle bir yaklaşım, şimdiki sözcük ile birlikte gizli katmanın önceki durumunu kullanarak bir sonraki sözcüğü tahmin ederken, gizli katmanda gelecek zaman adımları için yararlı olabilecek bilgilerin tutulmasını sağlamaz [55]. Bundan dolayı sinir ağının, gizli katmanda neyin saklanacağını öğrenmesini sağlayan, normal geri yayılım algoritmasını zaman adımları boyunca uygulayan BPTT (Back-Propagation Through Time) [56] algoritması kullanılır.

BPTT algoritması, zaman adımları boyunca ağın yinelemeli ağırlıkları üzerinde hatanın türevlerini geriye doğru yaymak için kullanılmaktadır. Böylece model, gizli katmanın yararlı durum bilgisini elde etmek üzere eğitilmektedir. Normal geri yayılım algoritması yinelemeli ağlarda görece daha düşük bir performans sergilemektedir [50]. BPTT algoritması yinelemeli sinir ağının çok katmanlı ileri beslemeli sinir ağı gibi işlem görmesini sağlar. Örneğin bir gizli katmana ve T zaman adımına sahip bir yinelemeli sinir ağı, T adet gizli katmana sahip derin ileri beslemeli sinir ağı gibi davranmaktadır. Böyle bir derin ileri beslemeli sinir ağı da normal gradyan düşürme (gradient descent) algoritmasıyla eğitilebilir.
Genelde T için sabit bir değer seçilir ve hatanın türevi bu zaman adımını geçmeyecek şekilde geri yayılır. Bu durumda, bu değerden uzun bağımlılıklar sinir ağı tarafından modellenemez.

4.3.1. LSTM ile Dizi Sınıflandırma

LSTM'in giriş katmanına dizi halinde bir girdi verildiğinde önceki durum bilgileri kullanılarak çıkışa yönlendirilir. Girdinin son elemanı da bu şekilde işlendikten sonra sınıflandırma işlemine geçilir. Bunun için tam bağlı (fully connected) bir çıkış katmanı ve katmanın çıktılarını olasılıklara çevirmek için softmax fonksiyonu kullanılır. Bu yapıya

çoka-bir (many-to-one) haritalama denir (Şekil 4.9). Genelde duygu sınıflandırma, konuşmacı ve dil tanıma gibi girdinin sıralı şekilde işlenerek sınıflandırıldığı uygulamalarda kullanılmaktadır. LSTM'in dizi sınıflandırma için kullanılması 5.4 Kısmında anlatılmıştır.



Şekil 4.9 LSTM'in çoka-bir mimariyle dizi sınıflandırma için kullanılması

4.3.2. LSTM ile Dil Modelleme

Dil modeli bir simge dizisinin olasılığını hesaplamaya yarar. Bu simge dizisi karakter, sözcük, fonem veya başka bir şeyden oluşabilir. Bu tez çalışmasında simgeler; fonemler (Kısım 5.3) ve ayrık birimlerdir (Kısım 5.4).

$w_1, w_2, w_3, \dots, w_n$ simgelerinden oluşan bir dizinin olasılığı $P(w_1, w_2, w_3, \dots, w_n)$'dir. Genelde dil tanımada [57, 58] dil modellerinden yararlanma yoluna gidilmektedir. Örneğin konuşma tanımada simge dizisi sözcüklerden veya fonemlerden meydana gelir. Bir konuşma tanıma sisteminin tahmin ettiği sözcüklerden oluşan aday birkaç cümleden en olası cümlenin seçimi düşünüldüğünde, dil modeline göre olasılığı diğerlerinden yüksek olan cümle seçilir. Dil tanımada ise tanımak istenilen sayıda dilin dil modelleri çıkartılır ve başlangıçta dili bilinmeyen bir cümlenin bu dil modellerine göre olasılıkları hesaplanır. Olasılığı en yüksek üreten model, verilen cümlenin dilinin modelidir.

$P(w_1, w_2, w_3, \dots, w_n)$ olasılığını kestirmek için eğitim verisinden bir istatistik çıkartılır. Bu istatistiği çıkarmak için N-gram dil modelleri kullanılabilir. Örneğin 2-gram dil modelinde simge dizisinin olasılığı şöyle hesaplanır:

$$P(w_1, w_2, w_3, \dots, w_n) = P(w_2|w_1)P(w_3|w_2) \dots P(w_n|w_{n-1}) \quad (4.18)$$

Çarpılan olasılık değerlerinin her biri, simgelerin sayılarak birbirine oranlanması yoluyla bulunur:

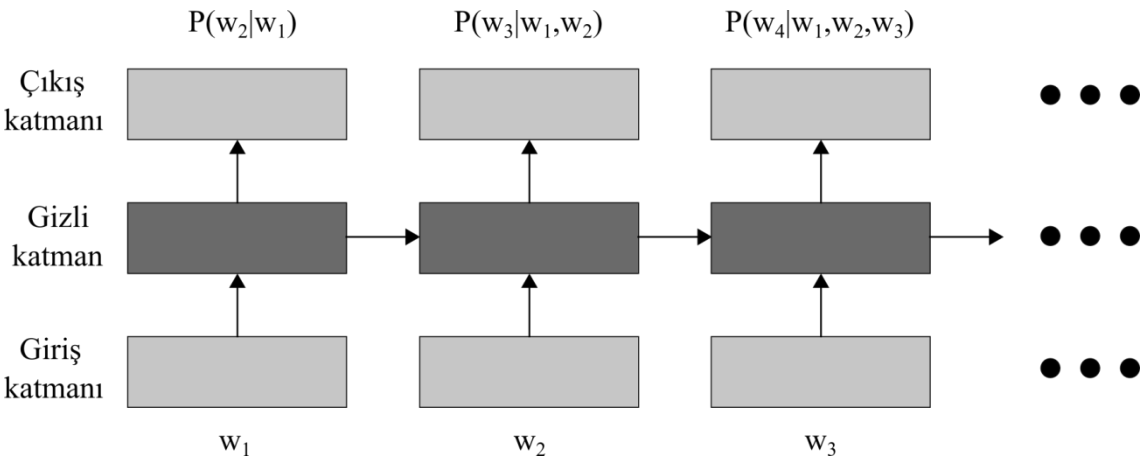
$$P(w_2|w_1) = \frac{say(w_1, w_2)}{say(w_1)} = \frac{\text{w}_1\text{ve } w_2\text{'nin birlikte bulunma sayısı}}{\text{w}_1 \text{ sayısı}} \quad (4.19)$$

N-gram dil modellerinin bütün simge kombinasyonlarına sahip olan büyük veri kümeleri üzerinde eğitilmesi gerekir. Yine de olmayan simge kombinasyonları için yumuşatma (smoothing) teknikleriyle [59] bu sorunun bir nebze üstesinden gelinmektedir. RNN dil modelleri n-gram dil modellerinin bu dezavantajını ortadan kaldırmakla beraber uzun dönemli geçmiş verisini de daha iyi modellemektedir [50, 60].

Yapay sinir ağlarının dil modellemede kullanıldığı ilk örneklerden birisi, Elman [51] tarafından önerilen, yinelemeli sinir ağlarının cümlelerin modellenmesi için kullanılmasıdır. Ardından doğal dillerin dil modelinin kurulması için ileri beslemeli sinir ağları kullanılmıştır [61]. Yinelemeli sinir ağlarının dil modelleme için kullanılmasıyla hem ileri beslemeli sinir ağıyla yapılan dil modellerine hem de n-gram modellerine göre daha iyi sonuçlar elde edildiği görülmüştür [50, 54].

Sinir ağlarıyla dil modellemede $P(b|a)$, sinir ağının bir sonraki simgeyi tahmin ederken ürettiği sonsal olasılık değeridir. Model bir zaman adımında simge tahmininde bulunurken o zamana kadar olan simgeleri hesaba katar. O halde LSTM dil modelinde simge dizisinin olasılığı şöyle hesaplanır:

$$P(w_1, w_2, w_3, \dots, w_n) = P(w_2|w_1)P(w_3|w_1, w_2) \dots P(w_n|w_1, w_2, \dots, w_{n-1}) \quad (4.20)$$

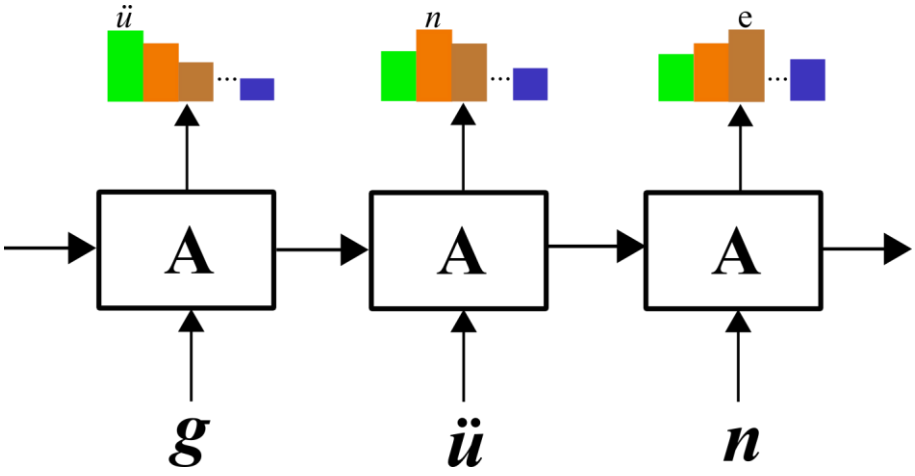


Şekil 4.10 LSTM dil modelinde olasılıkların her adımda hesaplanması.

Şekil 4.10'da girişe her bir zaman adımında genelde, bir simge için ilgili pozisyonda 1 diğerlerinde 0 olan bir *one-hot* vektör verilir. *One-hot* vektörün boyutu muhtemel simge sayısı kadardır. Hedef vektör de yine *one-hot* şeklinde kodlanmıştır. Ağın çıkış vektörü *one-hot* kodlamaya karşılık gelecek şekilde sonsal olasılıklardır. Çıkıştaki *one-hot* vektörde etiketi 1 olarak verilmiş girdi örneğinin olasılığı maksimize edilmeye, diğerlerinin ise minimize edilmeye çalışılır.

Şekil 4.10'da her bir zaman adımında elde edilen olasılık değerlerinin çarpılmasıyla verilen simge dizisinin olasılığı bulunmaktadır. Bununla birlikte, olasılıkları çarpmak yerine her birinin negatif logaritmalarını alıp toplayarak çapraz entropiler de bulunabilir. Çapraz entropi ne kadar düşükse dil modelinin tahminleri o kadar iyidir. Örneğin bir metinde, olasılığı 1 olan her karakteri tahmin eden bir dil modeli, sıfır çapraz entropi ile sonuçlanır. Diğer bir deyişle, var olan bir karaktere 0 olasılık değeri vermek yüksek maliyete neden olur. N adet karakter içeren metin için çapraz entropi şöyle hesaplanır:

$$CE = -\frac{1}{N}\sum_{i}^{N} \log_2 P(w_i|w_{i-1}, w_{i-2}, \dots, w_1) \tag{4.21}$$



Şekil 4.11 Örnek bir dizinin (“gün”) çıktılarının maksimize edilmesi.

Şekil 4.11’de "gün" dizisi için karakter bazlı dil modeli şöyle işler: Model, önceki karakterleri de hesaba katarak g karakterini gördüğünde örneğin % 40 olasılıkla ü karakterini tahmin ederse çapraz entropi $-\log_2(0{,}4) \cong 1{,}32$ olur. Model ü karakterini gördüğünde n karakterini örneğin % 55 olasılıkla tahmin ederse çapraz entropi $-\log_2(0{,}55) \cong 0{,}86$ olur. İki adımdaki ortalama çapraz entropi 1,09 olur: $P(\text{gün}) = -\frac{1}{2}\{\log_2 P(\text{ü}|\text{g}) + \log_2 P(\text{n}|\text{gü})\}$

Dil modellerinin ne kadar iyi tahminde bulunduğunu ölçmek için genelde perplexity (karışıklık) metriği kullanılır. Çapraz entropisi bilinen modelin karışıklığı şöyle hesaplanır:

$$PP = 2^{CE} \tag{4.22}$$

Burada istenen düşük entropi, dolayısıyla düşük karışıklık değeridir. LSTM dil modelinin fonem bazlı kullanılması Kısım 5.3'te, ayrık birim bazında kullanılması Kısım 5.4’te anlatılacaktır.

5. AĞIZ TANIMA

5.1. Genel Bilgiler

Bir konuşmanın veya metnin dilini otomatik olarak belirlemeye dil tanıma denir. Konuşma dilinin tanınması, bir konuşma örneğinde konuşulan dilin saptanması olarak tanımlanır [62]. Ağız tanıma ise, belli bir dilde verilen bir konuşmanın ağzının belirlenmesi olarak tanımlanmaktadır [63].

Konuşulan dillerin ayırt edilmesi insanlar için doğuştan gelen bir yetenektir [64]. Diğer yapay zekâ teknolojilerinde olduğu gibi konuşma dilinin tanınması da bu yeteneğin makineler tarafından taklit edilmesini amaçlamaktadır.

Dünyada binlerce konuşma dilinin olduğu tahmin edilmektedir [65]. İnsanlar, dilleri işitme sistemindeki algısal süreçlerden geçirmesi sonucunda tanımaktadır. Bu yüzden insanların kullandığı bu algısal ipuçları otomatik konuşma dili tanıma çalışmalarına esin kaynağı olmuştur [64]. Genelde bilgisayar ortamında yapılan çalışmaların çoğu metin üzerinden dil tanıma üzerine olmuştur. Sesli ifadeye göre farklılık gösteren metin-tabanlı dil tanıma yaklaşımı dilin sözcük veya sözcük-altı birimler gibi yazınsal özelliklerine dayanır. Metin-tabanlı dil tanıma probleminin önemli ölçüde üstesinden gelindiği söylenebilir [66]. Ancak konuşulan dilin tanınması tamamen farklı ve metin tabanlı dil tanımaya göre nispeten daha zordur.

İnsanlar üzerinde yapılan dinleme deneylerinde, dil sınıflarını belirlemek için geniş anlamda alt düzey (preleksikal) ve üst düzey (leksikal) olmak üzere iki ipucu kullanıldığı belirlenmiştir [64, 67]. Akustik, fonetik, fonotaktik, ritim ve tonlama özellikleri preleksikal bilgiyi oluştururlar [68]. Bunlar doğrudan konuşma sinyalinden elde edilebilen bilgilerdir. Sözcük anlamı ve söz dizimi ise leksikal bilgiyi temsil etmektedir. Bebeklerin, dilleri ayırt etmede preleksikal ipuçlarını başarılı olarak kullandıkları ortaya çıkarılmıştır [68]. Burada bebeklerin, leksikal bilgilerden ziyade, konuşmanın seslerine, ritmine, tonlamasına göre ayırt ettiği açıktır. Hiç bilmediği iki dilin ayırt edilmesi söz konusu olduğunda bir yetişkin de sadece preleksikal bilgisiyle karar vermektedir. Bebeğin dil deneyimi geliştikçe veya yetişkinin aşina olduğu diller arttıkça leksikal bilgi belirleyici olmaktadır [64].

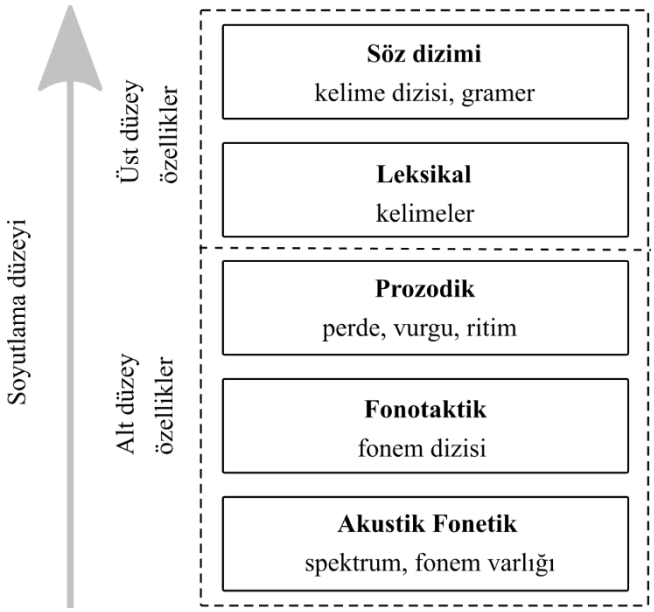
Ağız tanıma, daha genel olan dil tanımanın özel bir durumudur ve dil tanıma probleminden daha zordur. Çünkü ağızlar, dillere göre daha fazla ortak noktaya sahiptir ve diller arasındaki farklar daha belirgindir. Ağız, aynı kökten geldiği bir standart dilden belli oranda ayrılabilen yerel konuşma biçimi olarak tanımlanmaktadır [7]. Ağızlar, ait olduğu dilden ve diğer ağızlardan daha çok sessel (fonolojik), kimi zaman şekilsel (morfolojik), bazen söz varlığı (leksikal) ve söz dizimi (sentaks) bakımlarından farklılık gösterir [69].

Aynı dili konuşan insanların konuşmalarını birbirinden ayıran çeşitli özellikler bulunur. Cinsiyet, yaş, duygu durumu gibi özelliklerin yanında konuşmacının sahip olduğu ağız farklılıkları bunların başında gelmektedir. Büyük ölçekli, konuşmacıdan bağımsız otomatik konuşma tanıma sistemlerinin oluşturulabilmesi için ağız farklılıklarının ele alınması gerekir [57]. Böylece konuşmacının konuştuğu ağzın tanınması ve konuşma tanıma sisteminin buna göre ilgili ağız modeline anahtarlanması düşünülebilir.

Otomatik konuşma tanıma sistemleri genelde dilin standart olarak belirlenmiş ağızlarına göre yapılmaktadır. Türkçenin standart ağzı olarak İstanbul ağzı benimsenmiştir. Ancak Türkiye'nin çeşitli bölgelerinde konuşulan ağızlar İstanbul ağzından birçok yönüyle farklıdır. Bu farklılıklar işlenerek Türkçe konuşma tanıma sistemlerinin tanıma başarımları arttırılabilir.

Şekil 5.1'de dil bilimsel spektrum yer almaktadır. Spektrumdaki seviyelerden hangisinde işlem yapılırsa, kullanılan özellikler o seviyenin adıyla anılmaktadır. Bu tez çalışmasında bu spektrumun preleksikal olarak adlandırılan alt düzey özellikler kullanılarak ağız tanıma yapılmıştır. Dil tanıma için bu spektrumun bütün seviyelerinde işlem yapılabilirken ağız tanımada daha çok preleksikal seviyelerinde işlemlerin yapıldığı görülmektedir [57, 70].

Alt düzeydeki özelliklerin çeşidine bağlı olarak kullanılan sınıflandırıcılar da değişkenlik göstermektedir. Genelde zamanın şartlarına göre en iyi olduğu düşünülen sinir ağları [71], Saklı Markov Modeli (Hidden Markov Model, HMM) [72], Gauss Karışım Modeli (Gaussian Mixture Model, GMM) [57], Destek Vektör Makineleri (Support Vector Machine, SVM) [73] gibi sınıflandırıcılar dil ve ağız tanıma sistemlerinde kullanılmıştır.



Şekil 5.1 Dil ve ağız tanımada kullanılan özellikleri gösteren dil bilimsel spektrum.

İnsanın çıkarabileceği sesler çok fazladır. Konuşma sesleri (phone) diller üstüdür. Fonemler (phoneme) ise belli bir dile özgüdür. Bir dilde bulunan fonemler diğer dillerde de olabilir. Ancak bu ortak fonemlerin kullanılma sıklığı dilden dile değişmektedir [64]. Her dilin kendine özgü fonemleri olduğuna göre sadece fonemlere ve fonemlerin kullanılma sıklığına bakılarak diller birbirinden ayrılabilir. Bir fonem bir dilde nadiren kullanılırken diğer bir dilde bu foneme sıklıkla rastlanabilir. Fonemlerin fiziksel düzeydeki farklılıkları akustik seviyede incelenmektedir.

Akustik yaklaşım, fonemlerin sinyal düzeyindeki özellikleriyle ilgilenir. Bunun için sinyalden spektral özniteliklerin çıkartılması ve bunlar arasındaki farkların işlenmesi gerekir. Genelde sinyalden MFCC [74] veya SDC [75] gibi yöntemlerle öznitelikler çıkartılmaktadır.

Fonotaktik terimi, bir dilde izin verilen hece yapılarını belirleyen kısıtlar olarak tanımlanabilir. Diğer bir deyişle bir dildeki fonemlerin dizilimini o dilin fonotaktik kuralları belirlemektedir. Bir fonem birden fazla dilde bulunabilmesine rağmen, fonem dizilimleri ve bu dizilimin sıklığı dile özgüdür. Bir fonem dizilimi bir dilde nadirken, diğer bir dilde sıkça kullanılabilir. Bu özellik kullanılarak diller birbirinden ayrılabilmektedir.

Bir dilin fonem dizilimi bakımından farklarının ortaya konulması için dilin profili çıkartılmalıdır. Dil profili genelde n-gram modelleriyle [76] çıkartılmaktadır. Son yıllarda dil modellerinin öğrenilmesi için LSTM tipindeki yinelemeli sinir ağları kullanılmaktadır [50, 54]. Bu tez çalışmasında LSTM sinir ağlarıyla dil modelleri çıkartılmıştır.

Fonemler parçalı (segmental) sesbirim olarak adlandırılırken, prozodi için parçalar-üstü (supra-segmental) ifadesi kullanılmaktadır. Bir konuşmadaki vurgu, ritim, tonlama ve süre özelliklerinin tümüne birden prozodi denilmektedir [77]. Diller genel itibariyle prozodik özelliklerine göre ayrılabilmektedir.

Vurgu, ritim ve tonlama özellikleri algısal düzeydedir [68] ancak bunların fiziksel olarak ölçülmesi gerekir. Bunların fiziksel düzeydeki karşılıkları enerji eğrisi, seslendirme süresi ve perde eğrisinden elde edilmektedir [78, 79]. Enerji ve perde eğrisinden öznitelikler çıkartılarak bu eğrilerin öz bilgileri elde edilebilmektedir [78, 80]. Bununla birlikte bu eğrileri ayrık birimlere dönüştürme [57, 81, 82] gibi yöntemler de prozodi bilgisini elde etmek için kullanılmaktadır.

Dil/ağız tanıma sistemlerinden; cevap süresinin kısa olması, tanınması istenen dil/ağız sayısının artması durumunda performansın düşmemesi, kısa test cümlelerinde dahi iyi performans göstermesi, konuşmacı ve diğer etkenlere karşı sistemin dayanıklı olması gibi nitelikler beklenmektedir. Bunların yanında bu sistemlerin üstesinden gelmesi gereken sorunlar da vardır. Bu sorunların başında, sistemlerin eğitimini sağlamak için gerekli veri kümelerinin olmaması gelmektedir. Bir diğeri, kısa (örneğin 3 saniye) konuşma örneklerinde bu sistemlerin performansının düşmesidir. Ancak kısa örneklerde bu sistemlerin başarısı çok önemlidir [83].

Bu tez çalışmasında, yukarıda bahsedilen sorunları çözmek amacıyla Türkçenin ağızları veri kümesi adında bir derlem oluşturulmuştur. Bu veri kümesindeki konuşma örnekleri 2-3 saniye uzunluğundadır. Böylece ağız tanıma sisteminin kısa örneklerle eğitilmesi ve denenmesi sağlanmıştır. Kısa örneklerde tanıma performansı özellikle akustik ve fonotaktik düzeyde yüksektir.

5.1.1. Ağız Tanıma Sistemi

Ağız ve genel olarak dil tanıma sistemi iki aşamaya ayrılmaktadır. İlkin konuşma sinyalinden dil bilimsel seviyeye göre (akustik, fonotaktik ve prozodik) gerekli bilgilerin çıkartılması aşaması gelmektedir. İkinci aşamada ise çıkartılan bilgilerle eğitilen model veya modeller bulunmaktadır.

Burada her ağız (D adet ağız) için ayrı ayrı parametreler çıkartılır ve bu parametrelerden her ağız için modeller (D adet model) eğitilir. Bu modellerden test örneği için en yüksek olasılığı veren model, test örneğinin ağzı olarak seçilmektedir. Bu eğitim şekline üretici eğitim modeli denilmektedir. Dil modellerinin çıkartılması, diğer bir deyişle ağza özgü istatistiklerin elde edilmesi bu eğitim modeliyle yapılmaktadır. Bu şekilde eğitilen bir model sadece eğitildiği ağzın özelliklerini öğrenmektedir. Bu şekilde eğitim modeli 5.3 Kısmında LSTM ile dil modellerinin çıkartıldığı fonotaktik yaklaşımda ve 5.4 Kısmında ayrık birimlerin dil modellerinin çıkartıldığı prozodik yaklaşımda kullanılmıştır.

Konuşma sinyalinden üretilen öznitelikler X olarak tanımlansın ve bunlardan eğitilen ağız/dil modelleri de $\{\lambda_d | d = 1,2,\dots,D\}$ şeklinde gösterilsin. Bu durumda verilen bir X'e göre en yüksek sonsal olasılığı (posterior) üreten λ_d modeli şu şekilde seçilir:

$$\widehat{D} = \text{argmax}_d P(\lambda_d | X) \tag{5.1}$$

Bayes kuralı uygulanırsa, (5.1) eşitliği şöyle ifade edilebilir:

$$\widehat{D} = \text{argmax}_d \frac{P(X|\lambda_d)\, P(\lambda_d)}{P(X)} \tag{5.2}$$

Burada $P(\lambda_d)$ ağız modellerinin öncel olasılığını (prior) gösterir. Ağız/dil tanıma için, bir sesli ifadenin bir ağza/dile ait olma ihtimalinin genelde eşit olduğu varsayılır. Ayrıca $P(X)$ olasılığı da bütün ağızlarda aynı olduğuna göre, ağız tanıma işlemi olabilirlik (likelihood) şekline dönüşmektedir:

$$\widehat{D} = \text{argmax}_d P(X|\lambda_d) \tag{5.3}$$

Verilen bir konuşma örneğini üretmesi muhtemel olan ağzın bulunması demek, X'in meydana gelme olasılığının en yüksek olduğu ağız modelinin bulunması demektir. Sinir ağlarında, softmax fonksiyonuyla doğrudan sonsal olasılıklar elde edildiğinden olabilirlik (likelihood) hesaplarına ihtiyaç yoktur, ancak örneğin Gauss karışım modelleri olabilirlik hesabı için kullanılabilir.

Her bir ağız için ayrı model eğitmek yerine bu ağızlar arasındaki farkların öğretildiği eğitim şekilleri de vardır. Bu şekildeki eğitime ayırıcı eğitim modeli denir. Sinir ağlarının klasik eğitim şekli budur. Bu şekilde, bütün ağız özellikleri tek bir modelle öğrenilerek bu ağızlar arasındaki farklardan sonuca gidilmektedir. Verilen bir test örneği için model D adet (ağız sayısı kadar) olasılık hesaplar ve bu olasılıklardan en yükseğini test örneğinin ağzı olarak seçer. 5.2 Kısmında akustik düzeyde bütün ağızların özniteliklerinin çıkartılıp tek bir CNN mimarisi ve softmax ile sınıflandırılması ayırıcı eğitim şeklindedir.

5.1.2. Literatür

Bu kısımda dil ve ağız tanımaya; akustik-fonetik, fonotaktik ve prozodik açıdan yaklaşan çalışmalar özetlenmiştir.

5.1.2.1. Akustik-Fonetik Düzeyinde Çalışmalar

Leonard [84] tarafından, dillerde bulunan bazı önemli referans sayılabilecek sesler elle çıkartılmıştır. Beş dil üzerinde yapılan tanıma işleminde % 80 doğruluk oranı yakalanmıştır.

Cimarusti ve Ives [85] çeşitli öznitelik vektörlerinden elde ettiği, 100 boyutlu bir polinom sınıflandırıcı kullanarak sadece akustik özelliklere dayalı bir dil tanıma gerçekleştirmiş ve % 84 doğruluk oranına ulaşmıştır. Bu sonuç sadece akustik özelliklerin dil tanımada kullanılabileceğini göstermiştir.

Sugiyama [86] vektör kuantalama yöntemini dil tanımaya uyarlamıştır. Nakagawa ve ark. [72] ise vektör kuantalama, ayrık HMM ve sürekli HMM-GMM yöntemlerini karşılaştırmıştır. Dört dil üzerinde % 81,1 ile sürekli HMM-GMM yönteminin en iyi sonucu verdiği gözlenmiştir.

Muthusamy ve ark. [87] geniş fonetik sınıflar, spektral öznitelikler ve perde eğrisi özniteliklerini kullanarak on dil üzerinde çalışma yapmıştır. Fonetik düzeyde dillerin birbirinden daha iyi ayrılabildiğini göstermiştir.

Yan [88] akustik, fonotaktik ve prozodik düzeyde çalışmıştır. Dokuz dil için altı fonem tanıyıcıyı HMM ve dil modeli mimarisinde kullanmış ve % 91 doğruluk oranı elde etmiştir.

Wong ve Sridharan [89] GMM – Evrensel Arkaplan Modeli (Universal Background Model, UBM) yaklaşımını konuşmacı ve dil tanımada kullanmıştır. Torres-Carrasquillo ve ark. [75] kaydırmalı delta kepstral (Shifted Delta Cepstral, SDC) özniteliklerini GMM-UBM sisteminde kullanarak Çin lehçeleri üzerinde başarı elde etmiştir.

Biadsy [57] çalışmasında Arapçanın dört lehçesi için akustik özniteliklerin çıkartılmasını takiben GMM–UBM yaklaşımını kullanarak bu özniteliklerin istatistiklerini çıkartmıştır. Bunlara ek olarak Arapça standart ağzı için bağlam-bağımlı (context-dependent) fonem tanıyıcı ile fonem farklılıklarını ele almıştır.

Hanani ve ark. [90] İngilizce konuşmalardan bölgesel ağızları ve etnik grupları belirlemek için, akustik ve fonotaktik özellikleri elde etmişler, bunun için GMM ve Destek vektör makineleri kullanmışlardır.

5.1.2.2. Fonotaktik Düzeyinde Çalışmalar

House ve Neuburg [73] tarafından elle transkript edilmiş metinler kullanılarak dil tanıma yapılmıştır. Geniş fonetik sınıf etiketlerini modellemek için HMM yöntemi kullanılmış ve sekiz dil üzerinde tanıma yapılmıştır. Fonotaktik kısıtların dil tanımaya uygulandığı ilk çalışmalardan biridir.

Li ve Edwards [91] daha sonra Markov modelleriyle geniş fonetik sınıfların modellenmesi yaklaşımını gerçek konuşma verileri üzerinde kullanmışlar ve % 80 doğruluk oranı elde etmişlerdir.

Schultz ve ark. [92] dil tanıma için konuşma tanıma sistemi kullanmayı önermiştir. Fonem ve sözcük düzeyinde dil tanıma yapılmış ve sözcük düzeyinde 3-gram dil modeliyle % 84 doğruluk oranı yakalanmıştır.

Zissman [63] fonotaktik bilgiyi modellemek için PPRLM mimarisini geliştirmiştir. N-gram dil modellerini fonem dizilim istatistiğini çıkartmak için kullanarak dilleri sınıflandırmıştır.

Daha sonra bu yaklaşım, sadece bir fonem tanıyıcı kullanarak İspanyolcanın iki lehçesi üzerinde iyi sonuçlar vermiştir [93]. Navratil [67] ise akustik ve fonotaktik özellikleri birlikte kullanmıştır.
Biadsy ve ark. [94] fonotaktik bilgi için Zissman'ın [93] çalışmasını lehçelere uyarlayarak PRLM ve PPRLM mimarileriyle Arap lehçelerini sınıflandırmıştır. Bunun için dokuz fonem tanıyıcı sistemi kullanmış ve buradan elde edilen fonemleri kullanarak her bir lehçenin n-gram dil modellerini çıkartmıştır.

Soltau ve ark. [95] n-gram dil modellerini ağız tanımada, Soufifar ve ark. [96] ise n-gramlarla birlikte i-vektörlerin dil tanımada kullanılmasını göstermişlerdir.

5.1.2.3. Prozodi Düzeyinde Çalışmalar

Foil [97], ritim ve tonlama özelliklerinden yedi prozodik özellik çıkartmış ayrıca bunlara formant frekanslarını eklemiştir. Kümeleme algoritması kullanarak üç dil üzerinde % 64 oranına ulaşmıştır.

Goodman ve ark. [98], Foil'in çalışmasını öznitelik vektörüne başka öznitelikler ekleyerek genişletmiştir. Ayrıca Thyme-Gobbel ve Hutchins [99] yine Foil'in öznitelik vektörünü genişleterek 224 elemana çıkartmıştır. Segment üstü yapılar üzerinde çalışıldığı için elde edilen özniteliklerin gürültüye ve diğer bozulmalara karşı dayanıklı olduğu gösterilmiştir.

Cummins ve ark. [100] temel frekans ve enerji eğrisini çıkartarak doğrudan LSTM sinir ağlarında kullanmıştır. Verilen eğrilerin her bir çerçevesini teker teker sinir ağına vermiş ve sonuç skoru için son 0,5 saniyenin sonuçlarının ortalamasını bularak dilleri sınıflandırmıştır.

Adami ve Hermansky [81], temel frekans (F0) ve enerji eğrilerini çıkartarak cümleleri bu eğriler yardımıyla segmentlere ayırmıştır. Segment içinde kalan temel frekans ve enerji eğrilerini ayrık birimlere dönüştürerek dil tanıma ve konuşmacı tanımaya uygulamıştır. Böylece % 35 eşit hata oranı yakalanmıştır. Ayrık birimlere süre özelliği katıldığında bu oran % 30'a, fonetik tabanlı sistemle birleştirildiğinde (fusion) % 21,7'e düşmüştür.

Rouas [82] prozodik değişimleri dil ve ağızlar için modellemiştir. Adami ve Hermansky'nin [81] çalışmasına benzer şekilde hecelere ayırma işlemi yapıldıktan sonra

her cümleyi ünlü-ünsüz örüntülerini çıkartarak sözde-hecelere (pseudo-syllable) ayırmıştır. Bu sözde heceleri temel birim olarak ele alıp her birim için 18 etiket çıkartmış ve bunları n-gramlarla modellemiştir.

Mary ve Yegnanarayana [101] hecelerin başlangıç noktalarını bularak her bir cümleyi bu noktalardan segmentlere ayırmıştır. Üç heceden 21 öznitelik çıkartarak ileri beslemeli sinir ağında kullanmıştır. 12 dil için % 32'lik bir eşit hata oranı elde edilmiştir.

Biadsy [57], Adami ve Hermansky'nin [81] çalışmasını esas alıp Arap lehçelerinin perde, ritim ve süre özniteliklerini çıkartarak HMM ile modellemiştir. Yaklaşımlar içinde en düşük performansı prozodi yönteminde almıştır. Bu bakımdan prozodik ve fonotaktik yaklaşımların sonuçlarının birleştirilmesi yoluna gitmiştir.

Martinez ve ark. [78] prozodik bilgilerin özniteliklerini Legendre polinomlarıyla yakınsayarak i-vektörlerle modellemiştir. Geleneksel olarak kullanılan vurgu, ritim, tonlama özelliklerine formantları da eklemiş [102] ve F_1 ve F_2 formantlarının dilleri tanımada etkili olduğunu göstermiştir.

5.1.2.4. Derin Sinir Ağlarının Kullanıldığı Çalışmalar

Montavon [39] derin sinir ağlarını Fransızca, İngilizce ve Almanca dillerini tanımak için kullanmıştır. 5 sn'lik konuşma örneklerini iki boyutlu spektrogramlara çevirerek Konvolüsyonel Zaman Gecikmeli Sinir Ağlarında (Convolutional Time Delay Neural Networks, CNN-TDNN) öznitelik vektörü olarak kullanmıştır. Örnek boyunu geniş tutarak konuşmanın aynı zamanda fonotaktik ve prozodik bilgisini de yakalamayı amaçlamıştır.

DNN'ler yeni öznitelik gösterimleri üretmek için tıkanıklık (bottleneck) katmanları ile kullanılabileceği gibi her bir sınıf için sonsal olasılıkları da [103–105] elde etmek için kullanılabilir.

Sabit uzunluklu vektörler (i-vector) [106] esasında konuşmacı tanıma için geliştirilen bir yaklaşım olsa da son zamanlarda dil tanımada da [105, 107, 108] popüler hale gelmiştir. Genelde konuşma sinyalleri önce Mel frekans katsayıları (MFCC) [74] gibi öznitelikleri çıkartıldıktan sonra her bir örnek cümle sabit uzunluklu i-vektörlere dönüştürülmektedir. I-vektörlerde giriş verisi sağlamak üzere sonsal olasılıkların ya da tıkanıklık özniteliklerinin

elde edilmesi için derin sinir ağlarının kullanıldığı uygulamalar görülmektedir [103, 109, 110].

Son zamanlarda LSTM tipindeki yinelemeli sinir ağları dil tanıma uygulamalarında sıklıkla kullanılmaya başlamıştır [111, 112]. Bununla birlikte ileri beslemeli sinir ağlarıyla konvolüsyonel sinir ağlarının (CNN), sıralı öznitelikleri işlemek için kullanıldığı görülmektedir [113, 114].

Özel olarak derin sinir ağlarının Fin dilindeki yabancı aksanların tanınmasında kullanıldığı [115], bunların yanında tıkanıklık özniteliklerinin i-vektörlerde kullanıldığı [116], ayrıca akustik, fonotaktik gibi değişik seviyelerin birlikte işlendiği çalışmalar [117] mevcuttur. Arapça ağızların tanınması için derin sinir ağlarının [118] ve CNN ile iki yönlü LSTM ağlarının kullanılması söz konusudur [119].

5.1.3. Veri Kümeleri

Ağız bilimi alanındaki çalışmalar, veri kümelerinin yeterli olmaması ve analiz sürecinin zaman alması nedeniyle sınırlıdır. Bu tez çalışması kapsamında Türkçe ağız bilimi çalışmalarına da yardımcı olması beklenilen bir derlem [120] oluşturulmuştur. Ayrıca 5.2 Kısmında karşılaştırma amaçlı olarak TIMIT [121] veri kümesi kullanılmıştır.

Veri kümelerindeki konuşma örnekleri ortalaması sıfır (0), standart sapması bir (1) olacak şekilde Eş. 5.4'teki gibi normalize edilmiştir. Ortalama ve standart sapma bütün örnekler üzerinden hesaplanmıştır. X her bir konuşma örneğini göstermektedir:

$$X = \frac{X - \mathrm{ort}(X)}{\mathrm{std}(X)} \tag{5.4}$$

5.1.3.1. Türkçe Ağızlar Veri Kümesi

Bu tez çalışması için Türkiye'nin dört ağız yöresinden (Ankara, Alanya, Kıbrıs ve Trabzon) toplanmış olan ses kayıtları derlenmiş ve işlenebilir hale getirilmiştir. Toplanan kayıtlar metne dayalı değildir ve spontane gelişen konuşmalardan oluşmaktadır. Veri kümesinde eşit sayıda kadın ve erkekten kayıtlar vardır. Bunların eşit olması veri kümesinin cinsiyete bağımlı olmasını engellemek içindir.

Türkçenin ağız özelliklerinin ortaya konulması amacıyla dil bilimciler ilgili ağız bölgelerine giderek kayıtlar toplamıştır. Konuşmacılardan gelenek-göreneklerinden bahsetmesi veya bir hatırasını anlatması istenmiştir. Bu şekilde elde edilen kayıtlar doğal ve uzun olmaktadır. Konuşmanın yeteri kadar uzun olması içeriğin fonetik olarak zengin olmasını sağlar. Konuşmacılar ağız özelliklerini yansıtacağı düşünülen kişiler arasından seçilmiştir. Seçilen kişilerin bulunduğu yöreyi çok fazla terk etmemesi, yaşlı olması ve eğitim seviyesinin düşük olması gibi özelliklerine dikkat edilmiştir. Bu özelliklere sahip kişilerde ağza özgü seslere rastlama ihtimali yüksektir [122].

Bu çalışmada kullanılan veri kümesi, dil bilimcilerden elde edilen kayıtlar üzerinde düzenlemeler yapılarak oluşturulmuştur. Gürültü, örnekleme frekansı ve kanal sayısı farklılığı gibi etkenler ve sessizlik bölgeleri giderilmiştir. Veri kümesinde Ankara ağzı için 0,8 saat, Alanya için 0,7 saat, Kıbrıs için 0,65 saat ve Trabzon için 0,55 saat olmak üzere toplamda 2,7 saat veri bulunmaktadır. Her bir ağız yöresi için iki erkek iki kadın dört kişiden kayıt vardır. Kayıtların örnekleme frekansları SoX [123] yazılımı kullanılarak 16 Khz'e dönüştürülmüştür. Kayıtlar sözcük bazında Praat [124] programıyla etiketlenmiştir. Kayıtlar daha sonra 2-3 saniyelik cümlelere ayrılmıştır. Böylece her ağız yöresi için yaklaşık olarak 400 cümle elde edilmiştir. Veri kümesinin sayısal bilgileri aşağıdadır (Çizelge 5.1).

Çizelge 5.1 Veri kümesinin sayısal bilgileri.

	Toplam	Ankara	Alanya	Kıbrıs	Trabzon
Konuşmacı	16	4	4	4	4
Sesli ifade	1595	420	410	385	380
Sözcük	3620	935	909	891	885
Uzunluk (saat)	2,7	0,8	0,7	0,65	0,55

5.1.3.2. TIMIT Veri Kümesi

Bu veri kümesi konuşma tanıma uygulamalarında sıklıkla kullanılmaktadır. Amerika Birleşik Devletleri'nin (ABD) 8 ağız bölgesinden toplamda 630 kişinin 10 ayrı cümleyi seslendirmesiyle toplamda 6300 cümleden oluşmaktadır. Toplamda 3.1 saatlik konuşma

verisi metne dayalı olarak telefon üzerinden kaydedilmiş ve 16 KHz ile örneklenmiştir. Fonetik ve kelime düzeyinde etiketlenmiştir.

Veri kümesinin yukarıda bahsedilen 10 cümleden SA kodlu 2 cümlesi (toplamda 1260 cümle) ağız tanıma uygulamalarında kullanılmaktadır. Her cümle yaklaşık 3 saniye uzunluğundadır. Bu tez çalışmasında ABD'nin Northern (DR2), South Midland (DR4), Southern (DR5) ve Western (DR7) ağız bölgeleri kullanılmıştır. Bu ağız bölgeleri, konuşma örneklerinin diğerlerinden fazla olması ve sayıca birbirine yakın olması nedeniyle seçilmiştir. Veri kümesi belli bir okuma metnine dayandığı için cümle sonu vurguları belirsizdir, bu yüzden 5.2 Kısmında Türkçe ağızlar veri kümesi ile karşılaştırma yapmak amacıyla kullanılmıştır.

5.2. Akustik Açıdan Türkçe Ağızlarının Tanınması

Akustik bilgi, gerçeklemesi nispeten kolay olması ve iyi performans göstermesi nedeniyle dil ve konuşmacı tanıma sistemlerinde otuz yıldan fazla bir süredir kullanılmaktadır [125]. Akustik olarak ağız tanıma, ağızların spektral dağılım bakımından farklılık arz etmesi esasına dayanır. Türkçenin ağızları üzerinde çalışan dil bilimcilere göre, Türkçe ağızlar ünlü-ünsüz bağlamında birbirinden ayrılabilmektedir (Bölüm 3). O halde Türkçenin ağızları, spektral öznitelikleri bakımından birbirinden ayırt edilebilir.

Ağız tanıma problemi akustik olarak Eş. 5.5'teki gibi modellenebilir. $D = \{D_1, D_2, \dots, D_n\}$ sınıflandırılmak istenen ağız kümesi, $\vec{a} = \{\vec{a}_1, \vec{a}_2, \dots, \vec{a}_F\}$ konuşma örneğinin spektral bilgisini veren çerçeve tabanlı öznitelik vektörleri olsun. Burada amaç, verilen bir konuşma örneğinden elde edilen çerçeve tabanlı spektral öznitelik vektörleri kullanılarak en yüksek sonsal olasılığı (posterior) veren $\widehat{D}$ ağız sınıfını bulmaktır.

$$\widehat{D} = argmax_i P_{D_i}(D_i|\vec{a}) \tag{5.5}$$

Konuşma sinyalinin akustik açıdan incelenmesi için sinyalin fiziksel düzeydeki özelliklerinin incelenmesi gerekir. Çıkartılan sesler insanın ses yolu tarafından filtrelenir. Ses yolunun şekli kesin olarak belirlenebilirse, çıkan fonemin de gösterimi doğru olarak elde edilebilir. Bunun için ses yolu şeklinin kısa zamanlı güç spektrumundan çıkartılarak parametrik hale getirilmesi gerekir. Bu şekilde ham konuşma sinyalinin özet bilgisi parametre çıkartma teknikleriyle elde edilmektedir. Bu öz bilgi elde edilirken sinyalin önemli özelliklerinin korunmasına, önemsizlerin ise atılmasına dikkat edilir. Konuşma

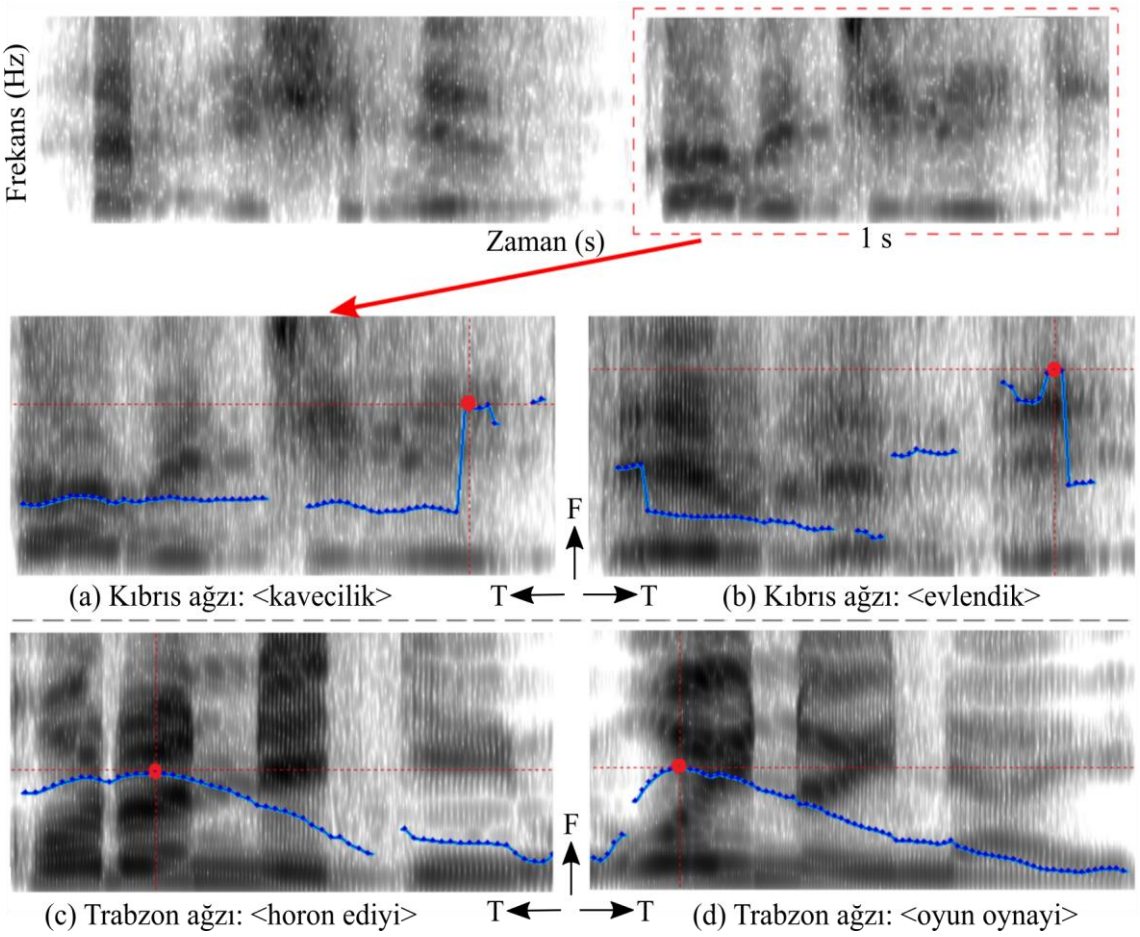
sinyalinin parametrik hale getirilmesiyle hem verinin boyutu düşürülmekte hem de asıl işi yapan sınıflandırıcıya ağızları birbirinden ayıracak en yararlı bilgi verilmektedir.

Birçok parametre çıkartma tekniği vardır. Bunlardan bazıları; Mel Frekans Kepstral Katsayıları (Mel Frequency Cepstral Coefficients, MFCC), Algısal Doğrusal Tahmin Katsayıları (Perceptual Linear Predictive coefficients, PLP), Kaydırılmış Delta Kepstrum (Shifted Delta Cepstrum, SDC) olarak sıralanabilir. Bu tekniklerle kepstrum katsayıları elde edildikten sonra sinyal içindeki zamansal bilgi de delta (Δ) ve delta-delta (ΔΔ) özellikleri alınarak kepstrum katsayılarına dahil edilebilir. Bununla birlikte, CNN türü sinir ağlarının son yıllarda popülerlik kazanmasıyla beraber akustik bilgiyi iyi temsil ettiği için logaritmik mel-spektrogramlar da kullanılmaktadır. Bu tez çalışmasında, öznitelikleri temsilen MFCC katsayıları ve log mel-spektrogramlar kullanılmıştır.

5.2.1. Tez Çalışmasının Akustik Açıdan Hipotezi

Türkçe veri kümesindeki ağız örnekleri üzerinde yapılan incelemelerde, genellikle cümle sonlarının daha vurgulu söylendiği belirlenmiştir. Cümle sonlarındaki tonlamaların şekli ağızdan ağza değişmektedir (Şekil 5.2).

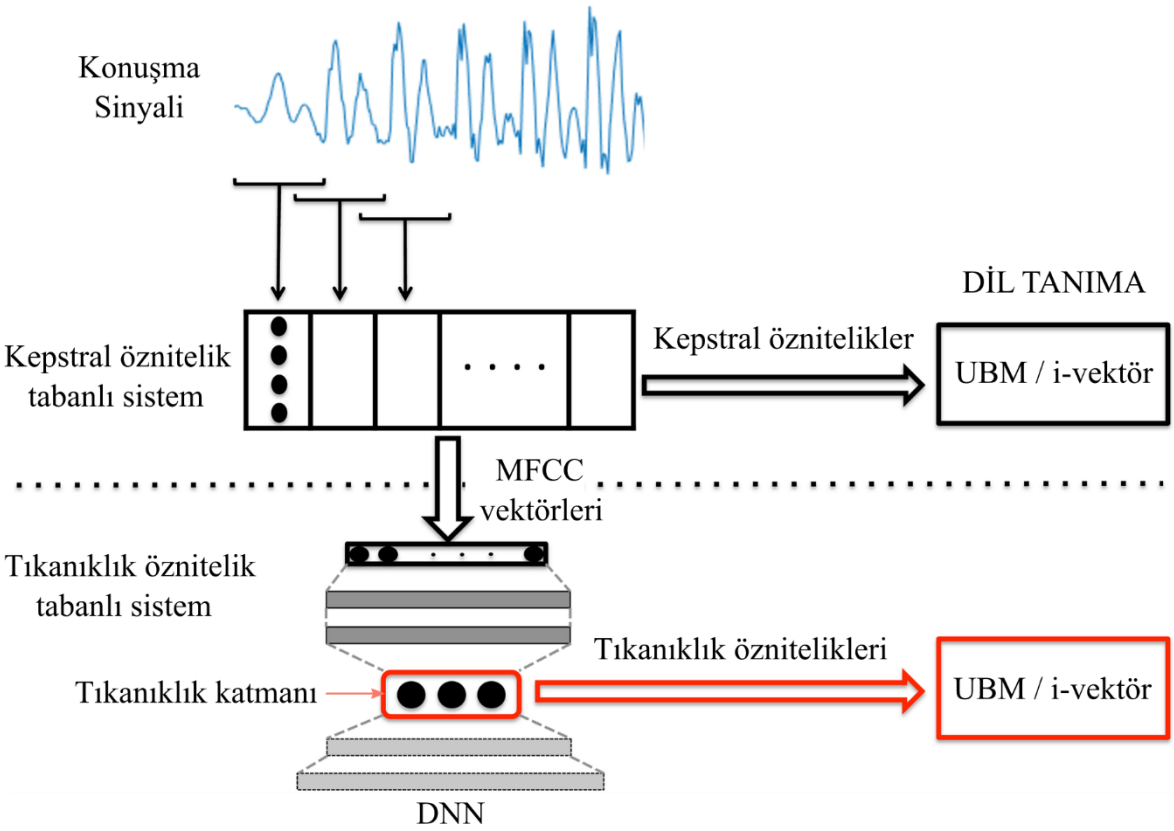
Şekil 5.2'de cümle sonundaki perde değişimleri (mavi eğri) ve perdenin zirve yaptığı, dolayısıyla en vurgulu söylenen kısımlar (kırmızı nokta) gösterilmiştir. Konuşma örneklerinde Kıbrıs ağzının perde frekansı genelde cümlenin sonuna doğru artış gösterirken, tam tersine Trabzon ağzında sönmektedir. Bu özellikten ötürü tez çalışmasının akustik açıdan hipotezi şudur: Sadece cümle sonundaki kısımlar kullanılarak Türkçenin ağızları akustik olarak birbirinden ayırt edilebilir.



Şekil 5.2 Cümle sonlarını gösteren spektrogram.

5.2.2. Derin Öğrenme ile Akustik Modelleme

Derin öğrenmenin ağız veya daha genel olan dil tanımaya akustik açıdan uygulanması genelde iki tür yaklaşımla yapılmaktadır. Bunların ilkinde, derin sinir ağları tıkanıklık (bottleneck) öznitelikleri oluşturacak şekilde fonem tanıyıcı olarak eğitilmektedir. Derin sinir ağında çıkış katmanına yakın olan bir gizli katman tıkanıklık katmanı olarak belirlenir. Genelde diğer gizli katmanlardan daha düşük boyutludur. Böylece girdi verisinin daha soyut ve daha düşük boyutlu hale getirilmiş bir üst görünümü elde edilmektedir. Bunun dil tanıma sistemlerine uygulanışı Şekil 5.3'te görülmektedir [126].



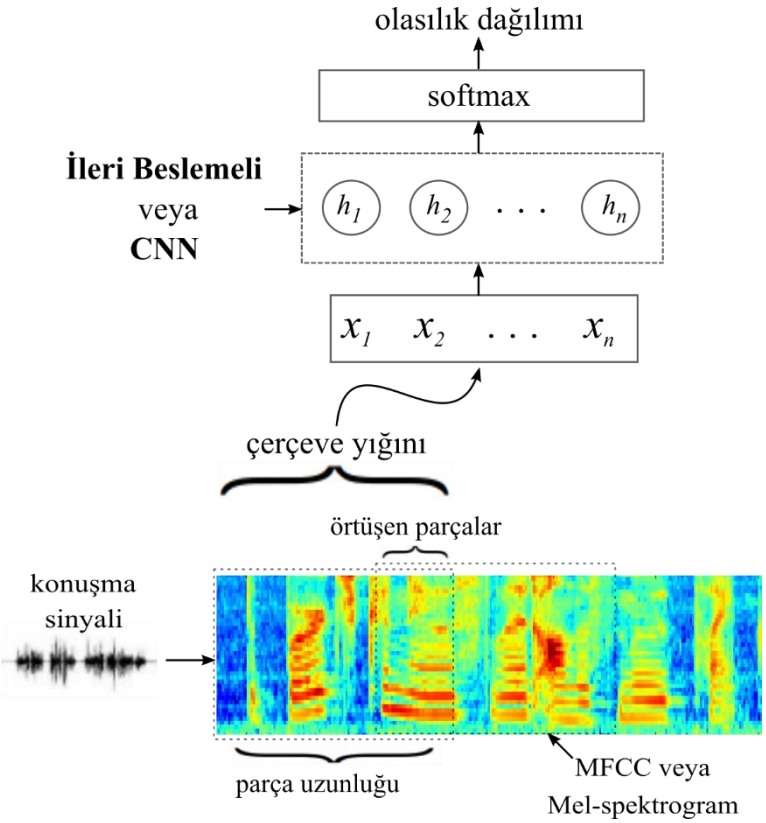
Şekil 5.3 Derin sinir ağlarında tıkanıklık katmanı ile özniteliklerin elde edilmesi.

Şekil 5.3'ten görüleceği gibi kepstral özniteliklere, derin sinir ağına verilerek tıkanıklık katmanında daha soyut bir üst görünüm kazandırılmaktadır. Burada DNN'ler hedef dillerden biri üzerinde konuşma tanıma maksatlı olarak eğitilmektedir. Bu yüzden softmax katmanının çıkışında elde edilen, konuşmayı oluşturan fonemlerin sonsal olasılıklarıdır. Ancak DNN burada öznitelik çıkartıcı olarak kullanıldığı için son katmandaki softmax sınıflandırıcının ürettiği sonsal olasılıklar ile ilgilenilmez. Tıkanıklık katmanından elde edilen parametreler daha sonra UBM ve i-vektör olarak adlandırılan başka bir sınıflandırma mekanizmasında girdi verisi olarak kullanılmaktadır. Bu şekilde, derin sinir ağlarının dil tanıma sistemlerine dolaylı olarak uygulandığı söylenebilir [104].

Derin sinir ağlarının akustik modellemeye uygulanmasında kullanılan ikinci yaklaşımda [39, 103], sinir ağı hem öznitelik çıkartıcı hem de sınıflandırıcı konumundadır. Sinir ağı, kullanılan birçok soyutlama katmanı ile birlikte konuşma özniteliklerinin değişik görünümlerini öğrenmek üzere eğitilmektedir. Böylece sinir ağının son katmanında veriye ilişkin sonsal olasılıklar hesaplanmaktadır.

Bu tez çalışmasında, akustik açıdan ağız tanıma için derin sinir ağları ikinci yaklaşımla ele alınmıştır. Bunun için çok katmanlı ileri beslemeli sinir ağı ve CNN mimarisi ağız tanımaya uygulanmıştır. Bu mimarilerin akustik modelleme için kullanılmasıyla gelişkin sistemler elde edildiği bilinmektedir [37, 39, 103, 127, 128].

Tez çalışmasında iki öznitelik gösterimi kullanılmıştır. Şekil 5.4'te konuşma örneklerinden elde edilen MFCC veya log mel-spektrogram özniteliklerinin derin sinir ağlarına giriş olarak verilmesi gösterilmiştir. Log mel-spektrogramlar CNN mimarisinde, MFCC ise geleneksel ileri beslemeli sinir ağında kullanılmak üzere iki ayrı uygulama yapılmıştır. Son aşamada kullanılan softmax sınıflandırıcıdan hedef ağız sınıflarının sonsal olasılık dağılımı elde edilmektedir.



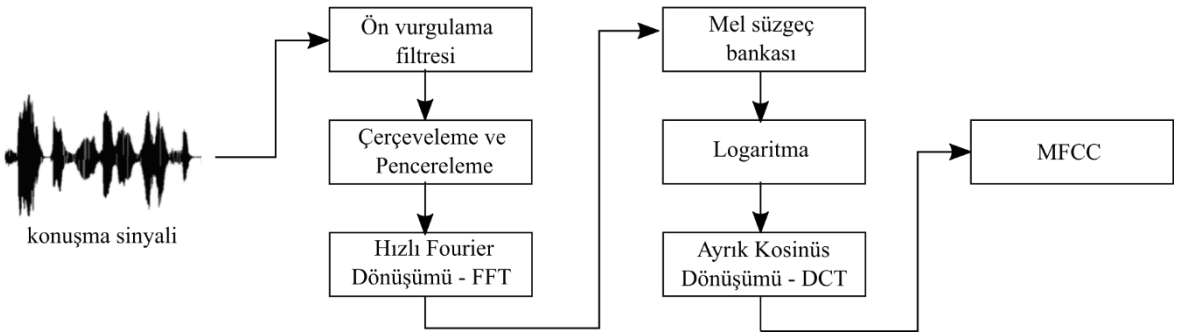
Şekil 5.4 Ağız tanımada derin sinir ağlarının doğrudan kullanılması.

5.2.3. Akustik Özniteliklerin Çıkartılması

5.2.3.1. MFCC

Konuşma, konuşmacı ve dil tanımada parametre çıkartma tekniği olarak MFCC [74] sıklıkla kullanılmaktadır. İnsan kulağı, sesi logaritmik olarak algılamaktadır. Örneğin 0-

1000 Hz frekans aralığı insan kulağında doğrusal olarak algılanır iken 1000 Hz'in üzerinde logaritmik algılanır. MFCC işlemiyle insan kulağının bu davranışı taklit edilir. Bununla birlikte MFCC parametreleri, konuşmanın dalga formunun sahip olduğu istenmeyen değişkenliklere daha az duyarlıdır.



Şekil 5.5 MFCC blok diyagramı.

Konuşma sinyali 0,9 ile 1,0 arasında değişen bir önvurgulama katsayısıyla çarpılarak yüksek frekanslarda bulunan enerji belirgin hale getirilir. Daha sonra konuşma sinyali, komşu çerçevelerin M adet örneği örtüşen olmak üzere, her biri N adet örnek içeren çerçevelere ayrılır. M ve N örnek sayısı genelde milisaniye cinsinden sırasıyla 10 ve 25 olarak seçilir. Daha sonra her çerçevenin başlangıç ve bitişindeki spektral bozulmaları minimize etmek için pencereleme adımı uygulanır. Genelde kullanılan Hamming pencereleme yöntemidir. Elde edilen sinyal hızlı Fourier dönüşümü (FFT, Fast Fourier Transform) ile zaman alanından frekans alanına dönüştürülür. Daha sonra, frekans spektrumuna K adet süzgeçten oluşan ve insan kulağının algılama şekline benzeyen Mel-ölçekli süzgeç bankası uygulanır. Mel-süzgeçleri birbiriyle örtüşen üçgensel pencereler şeklindedir ve bu pencerelerin 1 Khz'e kadar olan kısmı düzgün, 1 Khz'in üstündeki kısımları logaritmik olarak dağılmıştır. Böylece 1 KHz'e kadar doğrusal, 1 KHz'in üzerinde ise logaritmik olan Mel-ölçekli frekans spektrumu elde edilir. Logaritmik mel spektrumu tekrar zaman alanına çevrilerek mel frekansı kepstral katsayıları (MFCC) hesaplanır. Frekans spektrumunun kepstral gösterimi, verilen sinyalin yerel spektral özellikleri için iyi bir gösterim şeklidir. Mel spektrum katsayıları gerçel sayı olduğundan buna Ayrık Kosinüs Dönüşümü (DCT, Discrete Cosine Transform) uygulanarak zaman alanına çevrilir. Sonuçta elde edilen K (20 veya 40) adet katsayıdan genelde 12'si (2-13 arası) MFCC

katsayısı olarak kullanılır. K adet katsayının ilki çerçevenin enerjisini gösterir. Bu katsayının da eklenmesiyle 13 katsayı elde edilir.

Konuşmanın dinamik özellikleri, dillerin birbirinden ayırt edilmesini sağlayan önemli bir özelliktir. Bu özellikleri yakalamak için genelde delta ve delta-delta katsayıları MFCC katsayılarına eklenmektedir. Bunun için 13 MFCC katsayısının birinci (delta) ve ikinci (delta-delta) türevleri alınarak sırasıyla çerçeveler arasındaki hız ve ivmelenme özellikleri hesaba katılır. Bunun sonucunda 13 katsayı hız için, 13 katsayı da ivmelenme için olmak üzere toplamda 39 katsayı elde edilir. Bu şekilde konuşma sinyalinden MFCC katsayıları çıkartılarak sınıflandırma işleminin öznitelik vektörü oluşturulmuş olur. MFCC, literatürde genel kabul görmüş ve başarılı bir öznitelik çıkartım yöntemi olduğu için bu tez çalışmasında kullanılmıştır.

5.2.3.2. Logaritmik Mel-spektrogram

MFCC için anlatılan DCT adımı uygulanmazsa frekans alanında mel ölçekli güç spektrumu elde edilmiş olur. Zaman-frekans uzayında elde edilen mel-spektrogramda sinyalin ne zaman hangi frekansta olduğu tespit edilebilir. Mel ölçekli spektrogram gösteriminin normal spektrogramdan farkı, doğrusal olmayan süzgeçler uygulanarak elde edilmesidir. Ayrıca normal spektrogramdan daha düşük boyutludur. Mel ölçekli spektrogramın girdi verisi olarak kullanılması derin öğrenmede sıklıkla karşılaşılan bir durumdur [41, 42].

CNN mimarisi uzaysal bilgiden örüntüyü öğrenmede yeteneklidir. Bu yüzden, ağızlarla ilgili akustik bilgiyi gösteren log mel-spektrogramların matris yapısındaki örüntüsünü öğrenmesi için CNN kullanılabilir.

5.2.4. Akustik Modelleme Uygulamaları

Veri kümesi *k*-katlamalı çapraz doğrulama (k-fold cross validation) yöntemiyle eğitim ve test kümesi olarak ayrıldı. Burada $k = 10$ olarak seçildi. Böylece tüm veriler 10 kümeye ayrılarak bunların 9'u eğitim, 1'i de test için kullanıldı. Bu parçaların eğitim ve test olarak ayrılması işlemi 10 defa arka arkaya yapıldı ve bu 10 denemenin ortalaması alınarak sonuç skoru elde edildi. Böyle yapılmasının nedeni rastlantısallığı azaltarak sonuç skorunun

tutarlılığını sağlamak ve belli bir eğitim kümesinden bağımsız olmasını sağlamaktır. Performans ölçütü olarak doğruluk (accuracy) oranı kullanılmıştır.

Eğitim ölçütü olarak çapraz entropi kullanıldığı için, bütün örnekler one-hot vektörü şeklinde doğru sınıf için 1, diğerleri için 0 olarak kodlanmıştır. Böylece softmax fonksiyonunun doğru sınıfın olasılığını arttırması, diğerlerini ise azaltması sağlanmıştır. Çapraz entropi ve softmax hesaplamaları için Bölüm 4'e başvurulabilir. Ayrıca sinir ağı mimarilerinin kurulumu ve eğitimi için Keras kütüphanesi [129] kullanılmıştır.

5.2.4.1. İleri Beslemeli Sinir Ağı ile Akustik Model

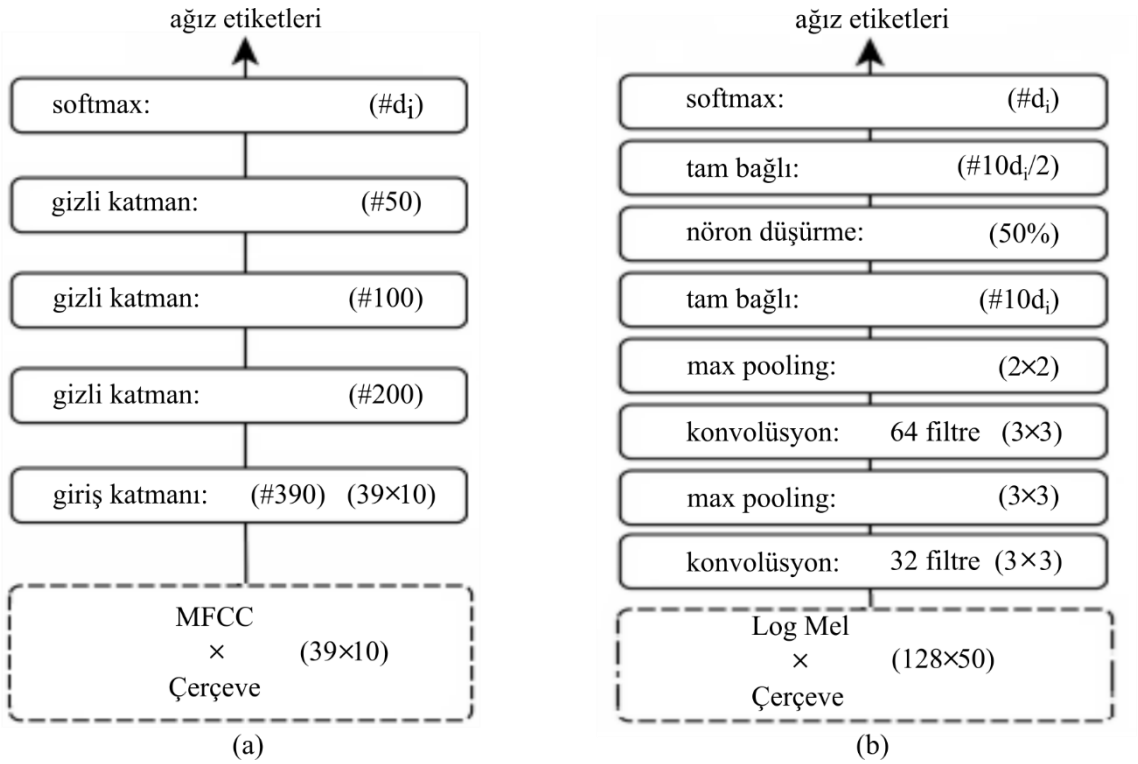
İleri beslemeli sinir ağları birden fazla işlem katmanıyla kendisine verilen giriş verisinin hiyerarşik temsilini çıkartarak son kararını vermektedir. Üç gizli katmanı olan ileri beslemeli sinir ağına giriş verisi olarak MFCC öznitelik vektörleri uygulanmıştır (MFCC-3K). Hipoteze uygun şekilde, cümle sonunu gösteren 0,5 ve 1 saniyelik kısımlardan sırasıyla 50 ve 100 çerçeve oluşturuldu. Her çerçeve için 39-boyutlu MFCC vektörü elde edildi. 10 çerçevelik parçalar halinde birbirine eklenerek sinir ağının girişine verildi (39×10). Bir sonraki parça önceki parçanın yarısıyla örtüşmektedir. Böyle yapılarak eğitim kümesi daha da arttırılmaktadır. Parçaların 10 çerçeveden oluşmasının nedeni bağlam bilgisinin yakalanmak istenmesidir [40]. Sinir ağının gizli katmanlarında sigmoid (lojistik) aktivasyon fonksiyonu, çıkış katmanında ise olasılıkları elde etmek için softmax fonksiyonu kullanıldı. Eğitim çapraz entropi ölçütüne göre SGD algoritmasıyla yapıldı.

Parçalar halinde işlem yapıldığından bir anda bir parça işlenir ve sonuçta bir zaman adımı (t) için sonsal olasılıklar hesaplanır. Bu yüzden, sonuç skoru için, bir ağza ait bir cümlenin bütün parçaları için üretilen sonsal olasılıklar hesaba katılmalıdır. Cümlenin parçalarının birbirinden bağımsız oldukları varsayılırsa, her parça için üretilen sonsal olasılıklar çarpılarak sonuç skoru bulunabilir. Ancak bu çok küçük değerlere neden olacağından, bunun yerine logaritmaların toplamı alınarak ortalama bulunabilir:

$$S_d = \frac{1}{N}\sum_{t=1}^{N} \log p(D_d|x_t, \theta) \tag{5.6}$$

S_d, d ağzı için verilen test örneğinin skorunu göstermektedir. $p(D_d|x_t, \theta)$ ise d ağzı için t zamanında verilen x_t parçası ve sinir ağının θ parametrelerine karşılık olarak çıkışta elde

edilen sınıf olasılığıdır. Bütün zaman adımlarının (N parça sayısı) sınıf olasılıkları toplanıp ortalaması alınarak sonuç skoru elde edilir.



Şekil 5.6 MFCC-3K modeli (a) ve CNN (b) hiper-parametreleri.

MFCC-3K modelinde karşılaştırma yapmak için hem cümle sonları (cs), hem de cümlenin tamamı kullanıldı. Farklı parametrelerle sınaması yapılan ileri beslemeli sinir ağının en iyi sonucu sağlayan parametreleri Şekil 5.6a'da verilmiştir. Şekilde $\#d_i$ ağız sınıfı sayısıdır ve bu sayı 4'tür.

5.2.4.2. CNN ile Akustik Model

CNN'ler yerel öznitelikleri zaman ve frekans alanında çıkartabilme özelliğine sahiptir. Bu bakımdan CNN'lerin ağızların akustik olarak ayrılmasında kullanılması iyi bir seçenektir. Cümle sonlarındaki tonlamadan kaynaklanan akustik değişimi yakalamak için her cümlenin son 0,5 ve 1 saniyelik kısımları kullanılmıştır. Bu kısımlardan log mel-spektrogram öznitelikleri çıkartıldı.

Açık kaynaklı librosa kütüphanesi [130] kullanılarak her çerçeve için 128 mel özniteliği çıkartıldı. Her cümlenin son 0,5s ve 1s uzunluğundaki kısımlarından sırasıyla 128×50 ve

128×100 ebatlı log mel-spektrogram elde edildi ve CNN girişine verildi (Şekil 5.6b). CNN modelinde sırasıyla 32 ve 64 filtreden oluşan iki konvolüsyon katmanı kullanıldı. Bu katmanlardan sırasıyla 32 ve 64 adet öznitelik haritası elde edilmektedir. Log mel-spektrogram matrisine 3×3 boyunda filtreler uygulanarak her iki eksen (frekans ve zaman) boyunca da konvolüsyon işlemi yapıldı. Aynı şekilde *pooling* penceresi de öznitelik haritalarının her iki ekseni boyunca uygulandı. Öğrenme oranı olarak $\alpha = 0{,}001$ değeri seçildi. CNN modeli için en iyi sonucu veren parametreler Şekil 5.6b'de gösterilmektedir. Bu modelde karşılaştırma yapmak için hem cümle sonları (cs), hem de cümlenin tamamı kullanıldı. Cümlenin tamamı kullanıldıysa, cümle 0,5 saniyelik parçalar halinde CNN girişine verildi. Bu durumda bütün parçaların skor ortalaması alınarak sonuç skoru elde edildi (Eş. 5.6).

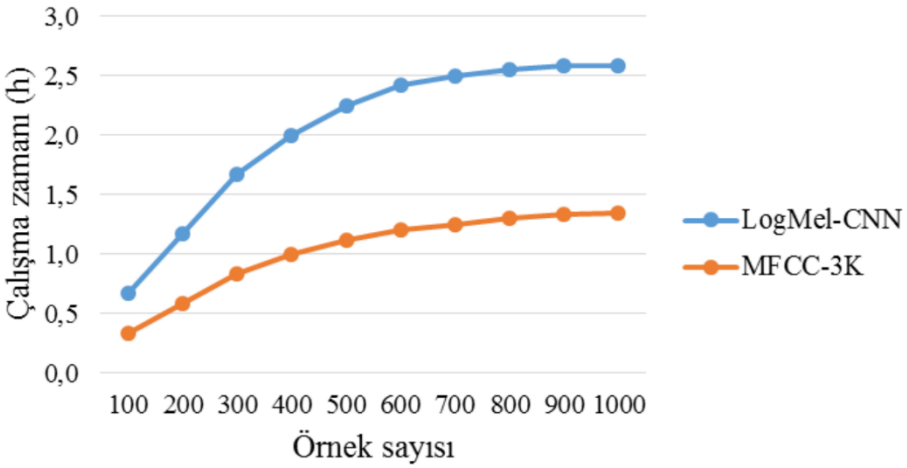
Şekil 5.6b'deki nöron düşürme adımı her parametre güncellemesinde, o katmanda bulunan % 50 oranında bağlantıyı rastgele koparark ağın aşırı öğrenmesinin önüne geçer ve boyutu düşürür. Ağda aktivasyon fonksiyonu olarak ReLU, çıkış katmanında sonsal olasılıklar için softmax fonksiyonu kullanıldı. Eğitim çapraz entropi ölçütüne göre SGD algoritmasıyla yapıldı.

5.2.4.3. TIMIT Veri Kümesi ile Karşılaştırma

Türkçenin ağızlarında gözlenen, cümle sonlarındaki tonlama farkından kaynaklanan ayrımın TIMIT veri kümesi için geçerli olup olmadığı araştırılmıştır. Türkçe belli bir metne dayalı olmayan spontane bir veri kümesi iken, TIMIT tam tersine metne dayalı bir veri kümesidir. Metne bağımlı veri kümelerinde tonlamalar fazla belirgin değildir. Bu iki özellikteki veri kümesinin karşılaştırılabilmesi için aynı şartların oluşturulması gerekir. Bu yüzden Türkçe veri kümesiyle aynı şekilde eğitim ve test kümeleri k-katlamalı ($k = 10$) çapraz doğrulama yöntemiyle ayrılmıştır. Sinir ağlarının hiper-parametreleri Türkçe veri kümesininkiyle aynıdır.

5.2.5. Bulgular ve Tartışmalar

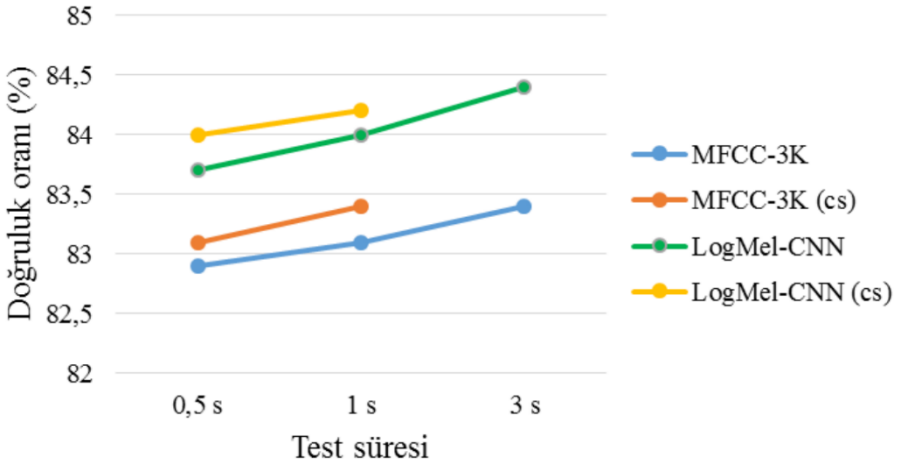
Uygulamalar Intel i7 işlemci yongasına sahip, 2,7 GHz frekanslı, 16 GB bellekli bir sistemde merkezi işlem birimi üzerinde yapılmıştır. İki uygulamanın veri kümesi üzerindeki çalışma zamanının örnek sayısına bağlı değişimi Şekil 5.7'de verilmiştir.



Şekil 5.7 Çalışma zamanı-örnek sayısı grafiği.

CNN modelinde görece daha büyük boyutlu log mel-spektrogram üzerinde gezdirilen filtrelerin (kernel) güncelleme işlemleri uzun sürmekte iken ileri beslemeli sinir ağı modelinin düşük çalışma süresine sahip olduğu görülmektedir.

Modeller Türkçe veri kümesinin 0,5 s, 1 s, ve 3 s süreli cümleleri üzerinde test edilmiş ve üretilen doğruluk oranları Şekil 5.8'de verilmiştir.



Şekil 5.8 Test süresine bağlı olarak uygulanan modellerin ürettiği doğruluk oranları.

MFCC-3K ve LogMel-CNN modelleri, hipotezde bahsedilen cümle sonu (cs) kısımlarının sınıflandırma performansına etkisini görmek için cümlenin tamamının kullanıldığı modelle

karşılaştırılmıştır. Burada cümle sonlarının süresi 0,5 s ve 1 s olarak belirlendiğinden 3 s süreli örnekler için deney yapılmamıştır.

Cümle sonu (cs) olarak gösterilen modeller, cümlenin tamamının kullanıldığı modellerden daha iyi sonuç vermektedir. Bu sonuç Türkçe sesli ifadelerin sadece cümle sonuna bakarak ağız bölgelerinin sınıflandırılabileceğini göstermektedir. Bu kadar kısa süreli konuşma örneklerinin bu oranları vermesi önemlidir. Çünkü sadece cümle sonlarının işlenmesi bilgisayar gücünden ve zamandan tasarruf anlamına gelmektedir. Ancak nispeten uzun (3 s) girdilerde cümlenin tamamının kullanıldığı modeller daha iyi performans göstermiştir. Bu da test süresi arttıkça doğruluk oranının arttığı görüşüne [110] uygundur.

TIMIT ile Türkçe veri kümesi üzerindeki sonuçların daha kolay karşılaştırılabilmesi için Şekil 5.8'deki Türkçe oranları aşağıda Çizelge 5.2'de birlikte verilmiştir.

Cümle sonu yaklaşımının kullanıldığı modellerin TIMIT veri kümesinde diğer modellerin gerisinde kaldığı gözlenmiştir. Bu sonuç, önerilen yaklaşımın metne dayalı olmayan, spontane veri kümeleri için daha uygun olduğunu göstermektedir. Çünkü metne dayalı konuşmalarda, spontane konuşmalarda olduğu şekliyle cümle tonlamalarındaki değişim çok belirgin değildir.

Çizelge 5.2 Türkçe ve TIMIT veri kümelerinin karşılaştırılması.

Veri kümesi ve süreye göre uygulanan modeller	Türkçe Ağızlar Veri Kümesi (%)			TIMIT Veri Kümesi (%)		
	0,5 s	1 s	3 s	0,5 s	1 s	3 s
MFCC-3K (cs)	83,1	**83,4**	-	81,2	81,3	-
MFCC-3K	82,9	83,1	**83,4**	82,7	82,9	**83,1**
LogMel-CNN (cs)	84,0	**84,2**	-	81,6	81,6	-
LogMel-CNN	83,7	84,0	**84,4**	83,0	83,2	**84,0**

Aşağıdaki karışıklık matrisi (confusion matrix) çizelgelerinde, kullanılan yöntemlerin Türkçe veri kümesi için ürettiği sınıflandırma başarımları verilmektedir. Çizelgelerde; satırlar hedef sınıfı (ground-truth), sütunlar ise modelin ürettiği sınıfı (output) gösterir. Çizelge 5.2'de en iyi sonucu veren modellerin Türkçenin ağızları üzerinde ürettiği

sonuçların karışıklık matrisleri sırasıyla aşağıdaki çizelgelerde görülmektedir. Bütün değerler yüzdeliktir.

Çizelge 5.3 MFCC-3K modellerinin karışıklık matrisi.

	Ankara	Alanya	Kıbrıs	Trabzon
Ankara	**82,2**	5,1	4,2	8,5
Alanya	5,4	**84,0**	6,2	4,4
Kıbrıs	6,2	6,5	**83,8**	3,5
Trabzon	6,5	5,4	4,6	**83,5**

Çizelge 5.4 LogMel-CNN modellerinin karışıklık matrisi.

	Ankara	Alanya	Kıbrıs	Trabzon
Ankara	**83,4**	4,2	5,4	7,0
Alanya	4,1	**85,4**	5,4	5,1
Kıbrıs	5,1	6,5	**84,8**	3,6
Trabzon	4,9	6,1	5,0	**84,0**

Karışıklık matrislerinde öne çıkan örüntü Alanya ve Kıbrıs ağızlarının birbiriyle karıştırılma olasılıklarının yüksek olmasıdır. Diğer bir örüntü ise Trabzon ağzıyla en az Kıbrıs ağzının karıştırılmasıdır. Bu iki sonuç coğrafi olarak yakın bölgelerin ağız özelliklerinin birbirine benzediği gerçeğini [11] desteklemektedir.

5.2.6. Akustik Açıdan Sonuçlar

Tez çalışmasının bu kısmında akustik modellemeyle Türkçenin ağızları üzerinde tanıma yapılmıştır. Tanıma için ileri beslemeli ve CNN mimarili sinir ağları kullanılmıştır. Bu ağlarda girdi verisi olarak MFCC ve log mel-spektrogram öznitelikleri kullanılmıştır.

Başlangıçta ortaya atılan "Cümle sonlarının tonlama farkından Türkçenin ağızları tanınabilir" hipotezi doğrulanmıştır. Bunun için konuşmalardan cümle sonlarının (0,5s ve 1s) kesilerek derin sinir ağlarında kullanılması yoluna gidilmiştir. Çizelge 5.2'de verilen

oranlar bu yaklaşımın ağız tanımada iyi sonuçlar verdiğini göstermiştir. Bu sonuç sadece cümle sonlarından elde edilen öznitelikler kullanılarak ağız bölgelerinin sınıflandırılabileceğini göstermektedir.

5.3. Fonotaktik Açıdan Türkçe Ağızlarının Tanınması

Fonotaktik, bir dilin/ağzın izin verilen fonem dizilimleri ile ilgilenir. Ağızlarda bulunan sesler ortak olabilir ancak bunlar dizilim bakımından farklılık gösterir. O halde ağızlardaki seslerin veya ses dizilerinin bulunma sıklığı (frekansı) karşılaştırılarak ağızlar birbirinden ayırt edilebilir. Bunun için öncelikle konuşmanın ses birimlerine ayrılması gerekir. Konuşmayı ses birimlerine ayırmak için klasik anlamda bir konuşma tanıyıcıya gereksinim vardır.

Yapılan işin gerçekte bir sınıflandırma problemi olduğu düşünülürse, ağız tanıma fonotaktik bakımından olasılıksal olarak Eş. 5.7'deki gibi modellenebilir. $D = \{D_1, D_2, \dots, D_n\}$ tanınması istenen ağız kümesi, $C = \{c_1, c_2, \dots, c_k\}$ fonemlerin bir dizisi olsun. Burada amaç, verilen bir konuşma örneğinden elde edilen fonem dizisi kullanılarak bu fonem dizisi için en yüksek sonsal olasılığı veren $\widehat{D}$ ağız sınıfını bulmaktır.

$$\widehat{D} = \text{argmax}_i P_{D_i}(D_i|C) \tag{5.7}$$

Türkçenin ağızları fonolojik, sözcüksel ve morfolojik açıdan birbirinden farklılık arz etmektedir. Bu yüzden seslerin ve ses dizilerinin sıklığı bakımından birbirinden ayırt edilebilir.

Standart fonotaktik yaklaşımları bir veya daha çok fonem tanıyıcıdan elde edilen fonem dizilerinin olasılıklarını çıkartır. Dil tanımada en fazla kullanılan fonotaktik yöntemi Paralel PRLM yöntemidir [68].

5.3.1. Paralel PRLM Yöntemi

Bir dilin fonotaktik kısıtlarını modellemede en iyi bilinen yöntem PRLM (Phoneme Recognition followed by Language Modelling, Fonem Tanıma ve Dil Modelleme) [93] yöntemidir. Bu yöntemde tanınması istenen ağız için konuşma örnekleri tek bir fonem tanıyıcı tarafından fonemlerine ayrılır. Daha sonra bu fonem dizileri üzerinden, tanınması istenen ağız için bir N-gram model eğitilir. Bu işlem bütün ağızlara uygulanır ve sonuçta

her ağız (n adet) için ayrı N-gram modeli elde edilerek bütün ağızların fonem dizilerinin olasılık dağılımları çıkartılmış olur. Tanıma işleminde ise, verilen konuşma örneği fonem tanıyıcıdan geçirilerek fonem dizilimi belirlenir. Bu fonem dizilimine, eğitimde elde edilmiş olan N-gram modelleri uygulanır. Bu modeller içinden en yüksek olasılığı üreten N-gram model, konuşma örneğinin ağız sınıfını verir. Bu yöntemde sadece bir fonem tanıyıcı olduğundan her ağız için bir olasılık değeri hesaplanır ve bunlardan yüksek olanı kolayca seçilebilir. Ancak birden fazla fonem tanıyıcı olması durumunda seçme işlemi zorlaşmaktadır.

Birden fazla sayıda (m) fonem tanıyıcının kullanılması durumunda bu mimari Paralel PRLM adını almaktadır. Bu yöntemde fonem tanıyıcılar paralel kullanılarak ağız konuşma örnekleri fonemlerine ayrılır. Daha sonra bu fonem dizileri üzerinden, tanınması istenen ağızların sayısı (n) kadar N-gram model eğitilir. Bu işlemin sonunda toplamda $m \times n$ adet N-gram model eğitilerek bütün ağızlardaki fonem dizilerinin olasılık dağılımları çıkartılmış olur. Tanıma işleminde ise verilen konuşma örneği fonem tanıyıcıdan geçirilir ve fonem dizilimi belirlenir. Bu fonem dizilimine eğitimde elde edilmiş olan N-gram modelleri uygulanır. Burada bir ağız için birden fazla (m) olasılık hesaplanmaktadır. Buradan hesaplanan olasılıklar birbiriyle çarpılarak hangi ağzın olasılığı yüksekse o seçilebilir. Fonem tanıyıcılar farklı olduğu için fonem dizilerinin birbirinden bağımsız oldukları varsayılırsa Eş. 5.7 bu sefer Eş. 5.8 şekline getirilebilir [57]:

$$\widehat{D} = \text{argmax}_i \; P(C_1; \lambda_i^1) \; P(C_2; \lambda_i^2) \; \dots \; P(C_m; \lambda_i^m) \tag{5.8}$$

Burada C_j, j fonem tanıyıcısından elde edilen fonem dizisini, λ_i^j ise i ağzının j. N-gram modelini göstermektedir. Basitçe Eş. 5.8'deki olasılıklar çarpılarak en yüksek sonucu veren ağzın seçilmesi yerine, N-gram modellerden elde edilen olasılıkların, destek vektör makineleri, sinir ağları veya lojistik regresyon gibi bir sınıflandırıcıya tabi tutulması düşünülebilir. Bu tez çalışmasında, dil modellerinden elde edilen skorlar için geleneksel sinir ağları kullanılmıştır.

Fonotaktik sistemde güvenilir dil modelleri oluşturulabilmesi için, kullanılan fonem tanıyıcıya büyük iş düşmektedir. Fonem tanıyıcıların geliştirilebilmesi için fonetik olarak yazıya geçirilmiş konuşma verilerinin olması gerekmektedir. Dil tanımada fonem tanıyıcının, tanınmak istenilen (hedef) dilde eğitilmesi gerekmez ancak hedef dildeki sesleri de kapsaması beklenir. Sadece İngilizce bir fonem tanıyıcı kullanarak dillerin

fonotaktik bilgisinin modellenmesi mümkün olmaktadır [131]. Ancak tek bir fonem tanıyıcı yerine birden fazla fonem tanıyıcı kullanılarak hedef dildeki seslerin yakalanmaya çalışılması daha iyi sonuçlar [57, 88, 93] vermektedir. Çünkü bir tanıyıcının tanımadığı bir fonemi, diğer tanıyıcı tanıyabilir. Ayrıca fonem tanıyıcının tutarlılık oranı önemlidir [132]. Tutarlılık her durumda aynı sonucu vermesi demektir. Örneğin, $a \rightarrow b$ gibi bir durumda a sesini her zaman b sesi olarak tanıyorsa bu sistem tutarlıdır [57].

Bu tez çalışmasında Paralel PRLM mimarisinde yer alan istatistiksel N-gram modeller yerine LSTM RNN dil modelleri kullanılmıştır. LSTM model N-gram gibi sabit ve küçük boyutlu geçmişi değil, değişken ve daha uzun dönemli geçmiş bilgisini hesaba katar. Ayrıca N-gram modelinin güvenilir olması için, fonemlerin bütün kombinasyonlarını içeren bir metin olması gerekir. Ancak LSTM ile yapılan dil modellerinde N-gramların bu dezavantajı ortadan kaldırılmakta ve her durumda bir olasılık elde edilebilmektedir.

5.3.2. Fonem Tanıyıcılar

Fonem tanıyıcı olarak Brno Üniversitesi tarafından geliştirilen PhnRec yazılımı [133] kullanılmıştır. Bu yazılımda İngilizce, Rusça, Macarca ve Çekçe dilleri üzerinde eğitilmiş dört adet fonem tanıyıcı bulunmaktadır. Bu fonem tanıyıcılar, sinir ağı ve Viterbi kod çözücü hibrit şekilde kullanılarak tasarlanmıştır.

Bu tez çalışmasında, bahsedilen tanıyıcılardan, daha tutarlı ve doğru sonuç verdiği gözlemlenen İngilizce ve Macarca (sırasıyla 39 ve 61 fonemden oluşmaktadır) fonem tanıyıcıları kullanılmıştır. Özellikle Macarcadaki <ü, ş, ç> gibi seslerin Türkçede de bulunması, bu fonem tanıyıcının seçilmesinde diğer etkendir. Örneğin <*giyerdik*> konuşma örneği bu tanıyıcılar tarafından fonemlerine şöyle ayrılmaktadır:

<giyerdik> → İngilizce fonem tanıyıcı → iy – eh – er – d – ih

<giyerdik> → Macarca fonem tanıyıcı → g – i – j – e: – d_: – i

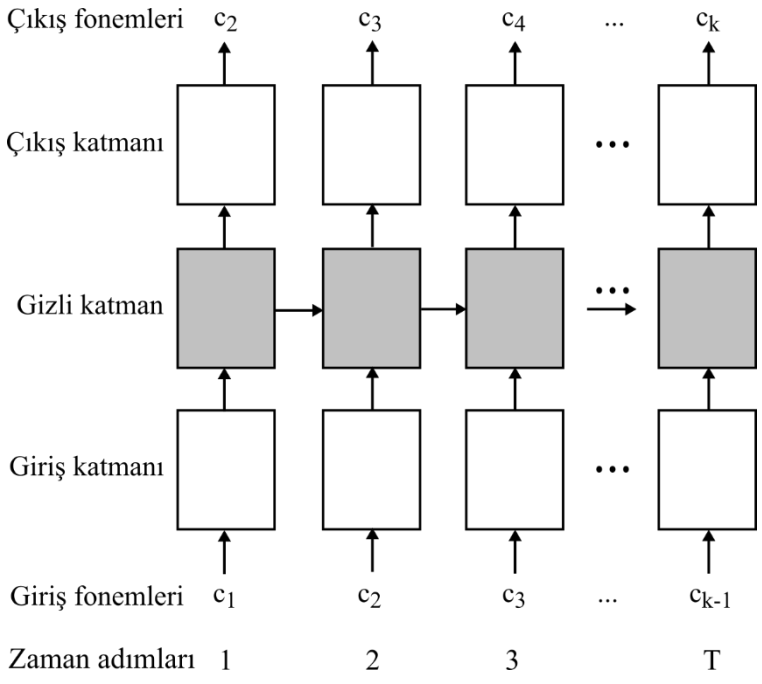
5.3.3. LSTM Sinir Ağları Dil Modeli ve PPRLM Mimarisi

LSTM sinir ağları, dil modelleme amacıyla kullanılmıştır. Burada amaç her ağız örneği için fonem tanıyıcılardan gelen fonem dizilerinin olasılığını LSTM ile modellemektir. LSTM ağın mevcut ve geçmiş fonem bilgisini kullanarak bir sonraki fonemi tahmin etmesi beklenir. Bu şekilde eğitilen $m \times n$ adet LSTM modelinden her biri, hangi ağız üzerinde

eğitilmişse o ağızda bulunan fonem dizisi için yüksek olasılık, o ağızda bulunmayan fonem dizisi için ise düşük olasılık vermektedir. Şekil 5.9'da bir ağzın dil modelinin LSTM sinir ağı ile eğitilmesi gösterilmiştir.

Fonem tanıyıcılardan, İngilizcede 39, Macarcada 61 fonem elde edilmektedir. Bu yüzden LSTM sinir ağının giriş ve çıkış vektörleri, İngilizce ve Macarca için sırasıyla 39 ve 61 boyutludur. Buna sözlük boyutu (B) denir. Giriş ve çıkış vektörleri one-hot şeklinde kodlanmıştır. Yani üretilen her fonem için sadece ilgili fonem pozisyonunda 1, diğerlerinde 0 bulunan bir seyrek matris yapısı vardır. Örneğin 4 ($B = 4$) fonemlik bir sözlük boyutu olsun. One-hot kodlamayla oluşan vektörler, birinci fonem $c_1 = [1,0,0,0]$, ikinci fonem $c_2 = [0,1,0,0]$ şeklinde devam ederek gösterilir.

LSTM sinir ağının çıkış katmanı giriş katmanı ile aynı boyuttadır. Çıkış katmanında sonsal olasılıkları elde etmek için softmax fonksiyonu kullanılmaktadır. Böylece çıkış katmanı mevcut fonem ile gizli katmanda depolanan bir önceki durum bilgisini birlikte hesaba katarak bir sonraki fonemin olasılık dağılımını çıkartır.

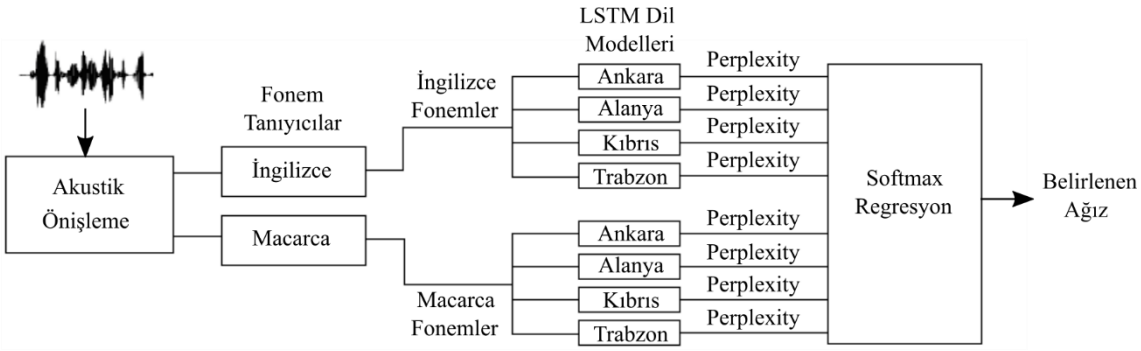


Şekil 5.9 Bir ağız için LSTM dil modelinin eğitilmesi.

Fonemlerin LSTM ağı içerisinde kullanımı Şekil 5.9'da gösterilmektedir. Burada giriş katmanına fonem tanıyıcıdan elde edilen fonemler verilmektedir. Her bir zaman adımında (T) sadece bir fonem işlenmekte ve ilgili örneğe ait fonem dizisinin tamamı işleninceye kadar süreç devam etmektedir. Gizli katmanın ürettiği sonuç hem çıkış katmanına gitmekte hem de sonraki durumun girişine verilmektedir. Aynı zaman adımında girişe verilen fonemden sonra gelen fonem, çıkış katmanının etiketlerini oluşturmaktadır. Böylece etiketi verilen bir sinir ağının gerçek ve tahmin edilen çıktısı arasında bir hata ölçümü yapılabilir. Bu şekilde eğitilen LSTM sinir ağıyla, her bir fonemden sonra gelen fonem tahmin edilerek ilgili ağzın dil modeli oluşturulur.

LSTM ağını eğitirken çapraz-entropi hata fonksiyonu kullanılmıştır. Bu hata fonksiyonu ağın ürettiği olasılığı maksimize etmeye yaramaktadır. Çapraz-entropi en düşük olduğunda eğitim sonlandırılır.

Her bir ağız için eğitilen LSTM dil modelleri Şekil 5.10'da gösterilen Paralel PRLM mimarisinde kullanılmıştır. Dört ağzın ($n = 4$) her biri için iki adet ($m = 2$) fonem tanıyıcı kullanıldığından toplamda 8 LSTM modeli vardır.



Şekil 5.10 Fonem tanıyıcılar ve LSTM dil modellerinden oluşan Paralel PRLM mimarisi.

Dil modellerinin başarımı *perplexity (PP)* denilen ve ağın ürettiği çapraz-entropi değeri ile doğru orantılı olan özel bir parametre ile ölçülür. İyi bir dil modeli, o dildeki bir konuşma örneğinden elde edilen bir fonem dizisine (cümle) düşük *PP* değeri vermelidir. Verilen bir cümle için her bir LSTM modelinin *perplexity* metriği hesaplanmıştır.

Dil modelleri eğitildikten sonra bu sefer konuşma örnekleri Şekil 5.10'daki gibi fonem tanıyıcılar tarafından fonem dizilerine dönüştürülür. Bu fonem dizileri ayrı ayrı LSTM dil modellerinden geçirilerek sonuçta 8 uzunluklu bir *perplexity* vektörü elde edilir. Bu yüzden softmax regresyon, girişinde 8, çıkışında 4 düğüm bulunan ileri beslemeli bir sinir ağıdır. Çıkış katmanında hedef sınıfı elde etmek için softmax fonksiyonu kullanılmaktadır. Böylece softmax regresyon, LSTM modellerinin ürettiği *PP* değerlerine göre hangi dil modelinin seçileceğine, dolayısıyla ağız sınıfına karar vermektedir. Doğru ağız sınıfının seçilmesi için LSTM ağında düşük *perplexity* üretilirken ileri beslemeli sinir ağında yüksek olasılık değeri üretilmelidir.

5.3.4. Uygulamalar

Fonem tabanlı dil modellerini karşılaştırmak için iki uygulama yapılmıştır. İlkinde N-gram dil modelleri, ikincisinde ise LSTM dil modelleri oluşturulmuştur. İki uygulamada dil modellerinden hesaplanan *perplexity* değerleri bir sonraki aşamada 8 girişli 4 çıkışlı ileri beslemeli sinir ağı sınıflandırıcısına verilmiştir. İleri beslemeli sinir ağının çıkışında softmax fonksiyonu ile olasılıklar hesaplanarak ağız sınıfı belirlenmiştir. Her iki uygulamada da doğruluk oranı ölçütü kullanılmıştır. Eğitim, test ve doğrulama kümeleri için konuşma örnekleri yüzde olarak 80, 10, 10 oranında ayrılmıştır. LSTM ağı kurulumu, ardından perplexity değerlerinin bulunması ve softmax regresyon hesaplamaları için Keras Kütüphanesinden [129] yararlanılmıştır.

5.3.4.1. N-gram Dil Modeli ile PPRLM

Veri kümesindeki bütün konuşma örnekleri yukarıda tanıtılan iki fonem tanıyıcıdan geçirilerek fonemlerine ayrıldı. Bu fonemler kullanılarak her ağız için ikişer olmak üzere 8 adet 3-gram dil modeli çıkartıldı. N-gram istatistiklerinin elde edilmesi için [134]'te verilen araç kullanıldı.

5.3.4.2. LSTM Dil Modeli ile PPRLM

Fonem tanıyıcılardan elde edilen fonemler kullanılarak her bir ağız için ikişer LSTM dil modeli eğitildi. Eğitilen dil modellerinin ürettiği *perplexity (PP)* değerleri hesaplandı. LSTM ağının eğitimi çapraz-entropi ölçütüne göre SGD algoritmasıyla yapıldı ve türevler, zaman boyunca hatanın geri yayılımı (BPTT) [56] algoritmasıyla hesaplandı. Başlangıçta ağırlık matrisleri sıfıra yakın değerlerle ilklendirilerek öğrenme hızı $\alpha = 0{,}1$ olarak

belirlendi. Her 10 örnekten sonra ağın öğrenme kabiliyeti, doğrulama verileriyle test edildi. Bu şekilde doğrulama verisinin olasılık değerlerinin artması durumunda eğitim devam etmekte, aksi halde α değeri yarıya düşürülmektedir. Olasılık değerinin önemli oranda artmaması halinde ise eğitim sonlandırılmaktadır. LSTM ağının giriş ve çıkış katmanları İngilizce tanıyıcıdan gelenler için 39, Macarca için 61 boyutludur. Yinelemeli olarak bağlanmış iki gizli katman ve her katmanda 50 LSTM birimi vardır. Zaman adımı $T = 10$ olarak alındı. Bu değer, açık hale getirilen sinir ağının katman sayısına denk gelmekte ve hafızada tutulacak fonem sayısını göstermektedir.

5.3.5. Bulgular ve Tartışma

Veri kümesinin 1s ve 3s uzunluğundaki sesli ifadeleri test edilmiş ve üretilen doğruluk oranları Çizelge 5.5'te verilmiştir.

LSTM dil modeli için farklı T zaman adımlarıyla denemeler yapılmıştır. $T = 10$ denemesinin daha iyi doğruluk oranı verdiği gözlenmiştir. Bu, zaman adımının yani geçmiş bilgisinin çok arttırılmasının elde edilen sonucu arttırmadığı anlamına gelebilir.

Çizelge 5.5 Uygulamaların test sürelerine bağlı doğruluk oranları.

Süreye göre uygulanan modeller	Doğruluk oranı (%)	
	1s	3s
PPRLM-3Gram	76,4	**79,0**
PPRLM-LSTM ($T = 10$)	84,2	**85,1**
PPRLM-LSTM ($T = 20$)	83,9	84,6

LSTM'in çok fazla geçmiş bilgisini gizli katman durumlarında tuttuğu bilinmektedir. Ancak bu kadar uzun geçmiş bilgisinin öğrenilmesinin bu çalışmada gerekli olmadığı düşünülerek küçük zaman adımlarıyla denemeler yapılmıştır. Zaman adımının çok fazla olması gizli katman sayısının da fazla olması anlamına gelmektedir. Fazla gizli katman sayısı hesaplama gücü ve bellek anlamında maliyetlere de neden olmaktadır.

LSTM sinir ağı oluşturulurken, LSTM katmanlarının üst üste yığılması yoluna gidilebilir. Bu durumda bir LSTM katmanının çıkışı hem aynı katmanda bir sonraki bellek hücresine hem de üst katmandaki LSTM bellek hücresine bağlanır. Bu yüzden farklı sayıda LSTM

katmanlarıyla da denemeler yapılmıştır. İki LSTM katmanının üst üste dizilmesi durumunda tek katmanlıdan daha iyi oranlar verdiği gözlendiği için çizelgede sadece bu modelin sonucuna yer verilmiştir.

LSTM ile oluşturulan dil modellerinin PPRLM mimarisinde kullanılması, 3-gram dil modellerinden daha iyi sonuçlar vermiştir. N-gram modellerin, bütün fonem kombinasyonlarını içeren metinlere ihtiyaç duyması gibi bazı dezavantajları olduğu bilinmektedir. LSTM ile bu dezavantajların üstesinden gelindiği Çizelge 5.5'teki sonuçlardan görülebilir. Her iki uygulamada da test süresi arttıkça sınıflandırma başarımı artmaktadır. Ancak N-gram modeli süreyle birlikte oransal olarak daha fazla artış göstermiştir.

Çizelge 5.5'te 4 ağız birden sınıflandırıldığı için ağızların birbirine göre durumları görülmemektedir. Bunu net olarak görmek için modellerin karışıklık matrislerine bakılabilir. Aşağıda LSTM'in en iyi modeli için verilen karışıklık matrisinde ağızların ikili sınıflandırma başarımları gösterilmiştir. Çizelgede; satırlar hedef sınıfı (ground-truth), sütunlar ise modelin tahmin ettiği sınıfı (output) gösterir.

Çizelge 5.6 PPRLM-LSTM ($T = 10$) modelinin 3s için karışıklık matrisi.

	Ankara	Alanya	Kıbrıs	Trabzon
Ankara	**84,4**	6,4	4,4	4,8
Alanya	4,7	**85,9**	5,2	4,2
Kıbrıs	4,8	6,6	**84,6**	4,0
Trabzon	6,1	4,8	3,6	**85,5**

Karışıklık matrisinde öne çıkan örüntülerden biri Alanya ve Kıbrıs ağızlarının birbiriyle karıştırılma olasılıklarının yüksek olmasıdır. Ayrıca Alanya ve Trabzon ağızları üzerinde, yöntem en iyi sonuçları vermiştir. Bu sonuç, bu iki ağzın kendilerine has özelliklerin diğerlerine göre fazla olmasına bağlanabilir. En düşük skorun Ankara ağzı için verildiği söylenebilir. Diğer bir örüntüye göre Trabzon ağzıyla en az Kıbrıs ağzı karıştırılmaktadır. Matriste en düşük oranlar bu iki ağız arasındadır.

5.3.6. Fonotaktik Açıdan Sonuçlar

Bu tez çalışmasında fonotaktik açıdan Türkçenin ağızları üzerinde tanıma işlemi yapılmıştır. Tanıma için derin öğrenme sinir ağları kullanılmıştır. Türkçenin ağızlarına yönelik, literatürde konuşma işleme ve makine öğrenmesi yöntemleriyle yapılmış bir sınıflandırma çalışması bulunmamaktadır. Bu yüzden başka çalışmaların sonuçlarıyla karşılaştırma yapma imkanı olmamıştır. Ayrıca tanıma için çok kısa sayılabilecek (1s ve 3s) konuşma örnekleri kullanılmış ancak çok fazla bilgi elde edilemediği için 0,5s süreli örnekler üzerinde deneme yapılmamıştır. Elde edilen sonuçlar tatmin edicidir.

Paralel PRLM, dil/ağız tanımada kullanılan popüler bir mimaridir. Bu mimaride yer alan dil modeli için genelde n-gram modelleri kullanılmaktadır. N-gram modeller sabit ve kısa geçmiş bilgisini modellediğinden, bu çalışmada n-gram yerine LSTM sinir ağları modelinin kullanılması önerilmektedir. Sonuçlardan görüldüğü kadarıyla ağız tanıma için LSTM dil modelinin PPRLM mimarisinde kullanılması iyi bir fikirdir.

Aik Lee ve ark. [135] tarafından 2016 yılında yapılan bir değerlendirme çalışmasına göre, fonotaktik için özellikle N-gram modellemenin ileriki yıllarda LSTM ağları kullanılarak yapılabileceği ileri sürülmektedir. Tez çalışmasının literatüre katkısı bu açıdan değerlendirilmektedir.

5.4. Prozodik Açıdan Türkçe Ağızlarının Tanınması

Ağızlar alt düzey ipuçları olan tonlama, ritim ve vurgu özelliklerine göre diğer ağızlardan ayrılmaktadır. Bu özelliklerin tümüne birden prozodi adı verilmektedir. Bu özellikler; temel frekans, süre ve enerji değerlerinin birleşiminden türetilen parametreler kullanılarak elde edilir.

Türkçenin ağızları prozodik özellikleri bakımından ayırt edilebilmektedir. Bu tez çalışmasının amaçlarından biri, Türkçenin ağızlarının prozodik özelliklerinin modellenerek otomatik sınıflandırılmasını sağlamaktır. Bunun için, ayırt edici prozodik özellikler, parametre çıkartma teknikleri, prozodik örüntünün çıkartılması ve modellenmesi, en iyi sonucu sağlayan yöntemlerin birleştirilerek daha iyi sonuçların elde edilmesi üzerinde durulmuştur.

Yapılan işin gerçekte bir sınıflandırma problemi olduğu düşünülürse, prozodik açıdan ağız tanıma olasılıksal olarak Eş. 5.9'daki gibi modellenebilir. $D = \{D_1, D_2, \dots, D_n\}$ tanınması istenen ağızlar kümesi olsun. $\vec{r} = \{\vec{r}_1, \vec{r}_2, \dots, \vec{r}_T\}$ bir cümlenin segmentlere ayrılmasıyla elde edilen temel frekans, enerji ve sürenin prozodik öznitelik vektörünü; $K = \{k_1, k_2, \dots, k_T\}$ ise cümlenin segmentlerinden elde edilen ayrık birim sözcüklerini göstersin. Burada amaç, verilen bir cümleden (utterance) çıkartılan $\vec{r}$ vektörü ve K sözcükleri kullanılarak bu cümle için en yüksek sonsal olasılığı veren $\widehat{D}$ ağız sınıfını bulmaktır.

$$\widehat{D} = \text{argmax}_i P_{D_i}(D_i|\vec{r}, K) \tag{5.9}$$

Prozodinin dil tanıma [80, 81, 102], konuşmacı tanıma [136, 137] gibi problemlerde kullanıldığı görülmektedir. Prozodik öznitelikler Legendre polinomlarıyla elde edilerek i-vektör yapılarıyla diller sınıflandırılmıştır [102]. Perde, enerji ve süre ölçümlerinin ayrık birimler haline getirilerek n-gramlarla modellenmesi ve böylece dillerin sınıflandırılması yapılmıştır [81, 82]. Arapça ağız tanıma için prozodik özniteliklerin kullanıldığı Gauss karışım modelleri ile sınıflandırma çalışmaları bulunmaktadır [57]. Bunların yanında prozodik özniteliklerin elle çıkartıldığı ve bunlarla dillerin sınıflandırıldığı çalışmalar mevcuttur [99, 138].

Geleneksel derin sinir ağları konuşmanın doğasından gelen uzun dönemli bağımlılıkları modelleyememektedir. Bu nedenle LSTM sinir ağları uzun dönemli bağlam bilgisini modellemeye daha uygundur [139]. Nitekim prozodik bilgi de böyle modellenebilecek niteliktedir. Bu tez çalışmasında perde ve enerji eğrileri Legendre polinomlarıyla parametrik hale getirilmiş ve polinom katsayıları öznitelik olarak kullanılmıştır. Bu öznitelikler LSTM katmanlı sinir ağıyla sınıflandırılmıştır. Ayrıca LSTM sinir ağları, dil modellerini çıkarmak için eğitilmiş, ayrık birimler olarak adlandırılan prozodik öznitelikleri modellemek ve ilgili ağzın profilini çıkarmak için kullanılmıştır.

5.4.1. Prozodik Öznitelikler

Konuşma olayı, bir dildeki anlamlı seslerin sıralı olarak bir araya getirilmesiyle meydana gelir. Ancak konuşma, yalnızca seslerin belli bir sırayla arka arkaya dizilmesi değil aynı zamanda doğallığı da içermelidir. Bazı özellikler konuşmayı doğal hale getirir. Konuşmayı doğal hale getiren özelliklerin tümüne prozodi adı verilmektedir. Perde değişimi, konuşmaya ayırt edilebilen melodik özellikler katar. Perdenin bu şekilde değişim göstermesiyle tonlama oluşur. Fonem ve hece düzeyindeki ses birimleri konuşmaya ritmik

özellikler katmak için kısaltılıp uzatılabilir. Ayrıca, konuşmada hece veya sözcükler, diğerlerine göre vurgulu söylenerek daha belirgin hale getirilebilir [101]. Bunların yanı sıra, tonlama, ritim ve vurgu gibi prozodik öznitelikler verilen mesajın anlaşılırlığını arttırır. Sayılan bu prozodik öznitelikler, algısal düzeydeki ipuçlarıdır ve fiziksel düzeyde temel frekans (f_0, perde), süre ve enerji parametrelerinin birleşimi ile ifade edilirler [99, 102].

Temel frekans, insan konuşma sesinin perde özelliğini verir ve bütün periyodik sinyallerde bulunur. Konuşma sinyalindeki perdenin zamansal dinamikleri tonlama ile ilgili bilgiler taşır. Konuşmanın belirli bir düzende gitmesini sağlayan ritmik karakteristiğiyle de alakalı bilgiler verir [140]. Enerji, konuşmanın ötümlü/ötümsüz bölgelerini tespit etmede kullanılır. Perde ve süre niteliğiyle birlikte kullanıldığında konuşmacının vurgu örgüsünü ortaya koyar. Konuşmada bulunan ünlü seslerin süresi ağza özgü bilgiler verir. Çünkü insanların bir sesi çıkartma süresi onların konuşma biçimini ve ritmini belirler, dolayısıyla ağızlar hakkında ipucu vermektedir.

Konuşmanın tonlama, ritim ve vurgu gibi prozodik öznitelikleri, konuşulan dilin kimliğine ilişkin bilgiler taşır. Konuşulan dilin belirlenmesi için yapılan dinleme deneylerinde, küçük çocukların tonlama ve ritim gibi öznitelikleri kullanarak karar verdikleri görülmüştür [68]. Aynı şekilde, yetişkinler de hiç aşina olmadığı diller söz konusu olduğunda, prozodik bilgilerine göre hareket etmektedir.

5.4.1.1. Prozodik Özniteliklerin Elde Edilmesi

Prozodik özniteliklerin çıkartılması için genelde 3 aşama söz konusudur [141]. Temel prozodi eğrilerinin (contour) çıkartılmasının ardından konuşma hece-benzeri birimlere ayrılır ve en sonunda bu birimler zamansal olarak modellenir. Perde ve enerji eğrilerinin çıkartılması için genelde otokorelasyon ve karekök ortalama (RMS, Root Mean Square) yöntemleri kullanılır.

5.4.1.1.1. Konuşmanın Hece-benzeri Birimlere Ayrılması

Konuşmanın heceler olarak arka arkaya dizilmesi, ağzın açılıp kapanması arasında bir ritmik değişime neden olmaktadır. Bu yüzden heceler prozodik olayların merkezindedir [101]. Diller geniş anlamda; vurgu zamanlı, hece zamanlı ve mora zamanlı olarak

ritmik/zamanlama özelliklerine göre ayrılmaktadır. Türkçe hece zamanlı diller grubundadır. Hece zamanlı dillerde ardışık hecelerin süresi yaklaşık olarak aynıdır. Bu nedenlerle bu çalışmada temel segment birimi olarak heceler kullanılmıştır.

Bir cümleyi hece veya hece-benzeri birimlere ayırma işlemine segmentasyon denilmektedir. Hecenin tam bir karşılığı olmasa da içinde bir ünlü sesin olduğu birimler hece olarak kabul edilir. Hece tanımı dilden dile değiştiğinden bu birimler hece veya hece-benzeri birim olarak adlandırılmaktadır [57, 82, 138]. Konuşmanın hece veya hece-benzeri birimlere ayrılması için genelde, konuşma tanıma sistemlerinin kullanıldığı ve kullanılmadığı yöntemler söz konusudur. Konuşma tanıma sisteminin kullanılmasıyla konuşmalar fonem ve hecelerine ayrılarak segmentasyon işlemi doğal olarak yapılmaktadır [82, 140]. Bu yöntem dile büyük ölçüde bağımlı olmasına karşın segmentlere ayırma konusunda başarılıdır. Hecelerin ünlü başlangıç noktalarının tespit edilerek ayrılması [101], konuşma tanıma sisteminin kullanılmadığı yöntemlere örnek olarak verilebilir. Ayrıca konuşma sinyalinden elde edilen enerji eğrisindeki vadiler kullanılarak hecelere ayırma işlemi yapılmaktadır [80, 136, 141]. Bunların yanı sıra akustik yöntemlerde uygulandığı şekliyle, sabit örtüşmeli pencereler kullanılarak da konuşmalar segmentlere ayrılmaktadır [102, 140].

5.4.1.1.2. Cümle Düzeyinde Modelleme

Prozodik özniteliklerin çıkartılması sürecinde üçüncü aşama modellemedir. Modelleme için eğri uydurma ve profil çıkarma olarak özetlenebilecek genelde iki yaklaşım mevcuttur. Eğri uydurma yaklaşımında, segmentlerine ayrılmış perde ve enerji eğrileri ayrık kosinüs dönüşümü ile modellenerek öznitelik vektörü elde edilir [141]. Bununla birlikte n. dereceden Legendre polinomlarıyla da her segmentten öznitelik vektörü çıkartılmaktadır [80, 102].

Profil çıkarma yaklaşımında, her örneğin dil modeli elde edilmektedir. Perde ve enerji eğrilerinden ayrık sınıflar adı verilen birimler oluşturulmakta ve bu birimler dilin profilini çıkarmak için kullanılmaktadır [82, 137]. Genelde dil modeli oluşturmak için istatistiksel n-gram modelinden faydalanılır.

İki yaklaşımın dışında, öznitelik vektörünün elle çıkartılması da söz konusu olmaktadır. [99] ve [97] perde ve enerji eğrilerinden istatistik yöntemlerle özniteliklerin çıkartıldığı ilk

çalışmalardandır. Ayrıca [138]'de elle çıkarılan 20 öznitelik prozodi temsili için kullanılmıştır. Cümle düzeyinde modelleme genelde iki yolla yapılmaktadır.

5.4.1.1.2.1. Legendre Polinomları ile Modelleme

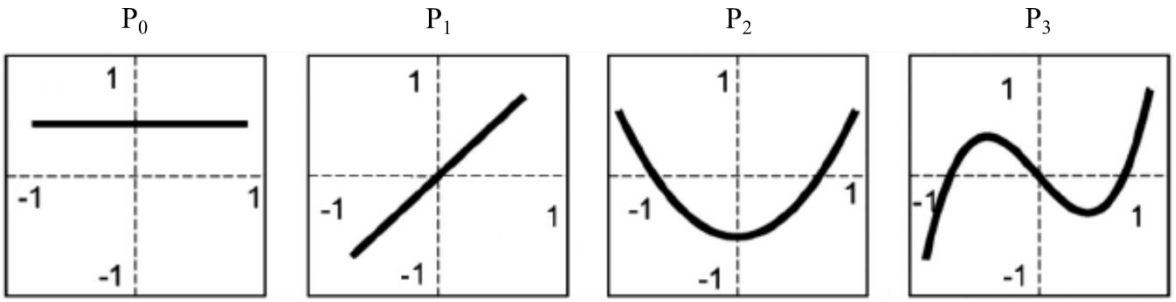
Dil tanımada, genelde konuşmanın kısa zamanlı kepstral öznitelikleri kullanılmaktadır. Bu özniteliklerin çıkartılması için en yaygın kullanılan yöntemlerden biri MFCC yöntemidir. Legendre polinom gösterimi, MFCC gibi, gerçek verinin daha düşük boyutlu ve daha öz halini sağlamaktadır. Akustik eğrilerin modellenmesinde Legendre polinomlarının kullanılması yaygındır.

Legendre polinomları ortogonal polinomlar sınıfında yer almaktadır. Ortogonal polinomlar, katsayılar arasındaki korelasyonları en aza indirgeme, yani büyüklüklerin birbirinden bağımsız olarak hesaplanması özelliğine sahiptir. Uyumdan (fitting) sonra her eğri bir modelle tanımlanmış olur. Bu model, Legendre polinomlarının toplamı olacak şekilde bir katsayılar kümesi (a_i) ile belirlenir. $f(x)$ formülü veya verisi bilinen bir fonksiyon, $P(x)$ ilgili fonksiyon noktalarından geçen model olsun.

$$f(x) = \sum_{i=0}^{N} a_i P_i(x) \tag{5.10}$$

Burada amaç $[-1,1]$ aralığında verilen bir $f(x)$ fonksiyonunu $n.$ dereceden $P(x)$ ile gösterilen Legendre polinomlar dizisi yardımıyla yakınsamaktır. Her bir polinomun a_i katsayısı bulunabilirse fonksiyon eğrisi üzerindeki noktalar yakınsanabilir. Fonksiyonu en iyi yakınsayan katsayıları elde etmek için en küçük kareler yöntemi ile hata miktarı bulunup düzeltme yapılmaktadır.

Polinomsal analizde, Fourier analizine benzer şekilde, düşük dereceli polinomlar eğrinin daha yavaş değişen özelliklerini gösterirken, yüksek dereceli polinomlar daha hızlı değişen özellikleri göstermektedir. Eğri ne kadar karmaşık olursa onu göstermek için o kadar çok polinom gerekir. İlk birkaç polinom (Şekil 5.11) fiziksel olarak yorumlanabilir: İlk polinom katsayısı a_0 ortalamayı, a_1 eğimi (artma-azalma), a_2 parabolü, a_3 ise eğrinin dalga şeklini ifade eder.



Şekil 5.11 İlk dört Legendre polinomu.

5.4.1.1.2.2. N-gram ile modelleme

Konuşma örnekleri segmentlere ayrıldıktan sonra perde (f_0) ve enerji eğrilerini tanımlayacak şekilde ayrık birimler elde edilir ve bunlar n-gramlarla modellenir [81, 82, 137]. Adami ve Hermansky [81], perde ve enerji eğrilerini hizaladıktan sonra bunların birbirine göre durumlarını tanımlayarak ayrık sınıfları oluşturur. Bu sayede 5 ayrık sınıf oluşmaktadır (Çizelge 5.7).

Aynı şekilde, süresi belli bir eşik değerinden kısa olan segmentler kısa (S), diğerleri uzun (L) olmak üzere iki sınıf belirlenir. Sonuçta, her bir cümle için, örneğin 5S 4S 2S 1S 3L 4L 5S 1S 2S 3S 4S 3S 2L 4S 5S gibi bir dizi ortaya çıkmaktadır. Bir dildeki bütün örnekler bu şekilde ayrık birimlerle ifade edildikten sonra her dilin n-gram modeli çıkartılır.

Bu çalışmada, konuşma örnekleri hece ortalarından segmentlere ayrıldıktan sonra her segmentteki ünlü (vowel) kimlikleri belirlenmiştir. Burada elde edilen 8 ünlü sınıfı da [137] çalışmasındaki ayrık sınıflarına eklenmiştir. Perde ve enerjinin birbirilerine göre artma ve azalma durumları P_1 Legendre polinomuyla elde edilmiştir.

Çizelge 5.7 Ayrık sınıflar ve etiketleri.

Birim etiketi	Perde ve Enerji Durumu
1	Artan Perde, Artan Enerji
2	Artan Perde, Azalan Enerji
3	Azalan Perde, Artan Enerji
4	Azalan Perde, Azalan Enerji
5	Ötümsüz (unvoiced) segment

Uzun dönemli geçmiş bilgisini başarılı şekilde öğrenmesi nedeniyle LSTM sinir ağı prozodi modellemede doğal seçenek haline gelmektedir. Bu yüzden prozodik segmentlerden hesaplanan Legendre polinom katsayıları öznitelik olarak LSTM sinir ağında kullanılmıştır. LSTM'in perde ve enerji verisini kullanarak dil tanıma problemine uygulandığı bir çalışma [100] vardır. Ancak buradaki fark, segment içindeki perde ve enerji verisinin doğrudan değil Legendre katsayılarına çevrilerek sinir ağına verilmesidir. LSTM ağı Legendre özniteliklerini kullanarak ağız tanıma için çoka-bir (many-to-one) haritalama yapmaktadır.

İkinci olarak, ağızların dil modeli (ağız profili veya fonolojik örüntüsü) LSTM sinir ağları ile oluşturulmuştur. Verilen bir konuşma örneği için en yüksek olasılığı üreten dil modeli, o örneğin ağız sınıfı olarak seçilmektedir.

5.4.2. Uygulamalar

5.4.2.1. Perde ve Enerji Eğrilerinin Çıkartılması

Perde ve enerji eğrilerinin elde edilmesi için Praat [124] programının varsayılan değerleri kullanıldı. Otokorelasyon yöntemine dayalı perde izleme algoritması yardımıyla perde eğrisi, yoğunluk izleme algoritmasıyla da enerji eğrisi çıkartıldı. Enerji eğrisi, sinyalin her çerçevesine karekök ortalama yöntemi uygulanarak elde edilir. Bunların oluşturduğu en iyi eğri yolunu bulmak içinse Viterbi algoritması kullanılır.

Normalizasyon işlemi için [102] çalışmasında anlatılan yol izlendi. Perde ve enerji değerlerinin logaritması alınarak insan algılama düzeyine getirildi. Enerji değerleri maksimum değerin çıkartılmasıyla, perde değerleri de ortalamanın çıkartılıp standart sapmaya bölünerek normalize edildi. Böyle yapılmasının nedeni, konuşmacılardan kaynaklanan istenmeyen değişkenliklerin azaltılmasıdır.

5.4.2.2. Segmentlere Ayırma

Türkçe ağızları veri kümesinde bulunan her bir ağız bölgesinden rasgele yaklaşık 50'şer cümle seçildi ve ünlü sesler kullanılarak bu cümlelerdeki hece ortaları elle tespit edildi. Daha sonra SpeechRate betiği kullanılarak [143] çalışmasında verilen referans değerlerle otomatik olarak hece ortaları bulundu. SpeechRate betiği Praat programı için yazılmıştır.

SpeechRate hece ortalarını bulurken enerji kontöründeki zirve noktalarının potansiyel hece ortası olduğunu varsayar ve ötümsüz bölgelerdeki zirve noktalarını dikkate almaz.

Otomatik olarak ve elle tespit edilen hece sayıları karşılaştırıldı ve aralarında $r = 0{,}86$ korelasyon hesaplandı. Ayrıca betik tarafından hece ortalarının bulunmasının doğruluk oranı % 85 olarak hesaplandı. Sonuçta elde edilen yüksek korelasyon ve doğruluk oranı Türkçe veri kümesinde bu betiğin hece ortalarını bulmak için kullanılabileceğini göstermiştir.

Hece ortaları bulunarak cümleler ham segmentlere ayrıldı. Ardından, sessiz bölgelerde perde eğrisinin (kontör) tanımsız olmasından yararlanarak cümle başı ve sonu tespit edildi. Cümle içinde, perdenin tanımsız olduğu (ötümsüz) bölgeler ise hesaba katılmadı. Böylece cümleler, üzerinde işlem yapılacak olan gerçek segmentlere ayrılmış oldu. Şekil 5.12'de örnek bir cümleden elde edilen perde ve enerji kontörleri ve bulunan segmentler alt alta hizalanmış şekilde görülmektedir.

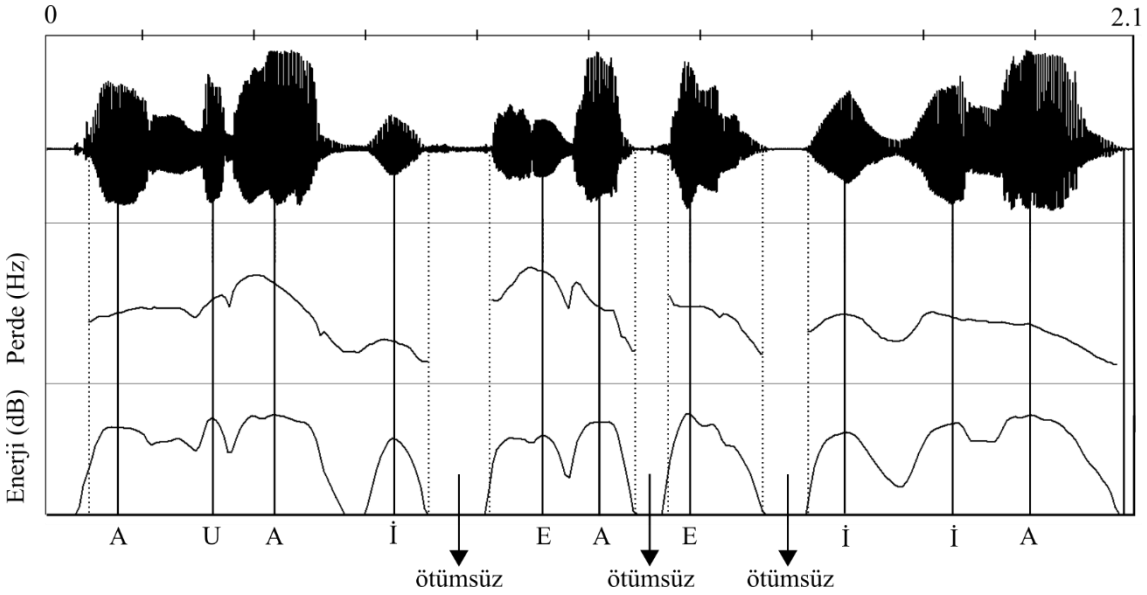
5.4.2.3. Modelleme

Cümleler hece ortalarından segmentlere ayrıldıktan sonra modelleme aşaması gelmektedir.

5.4.2.3.1. Legendre Polinomları ile Modelleme

Her segmentteki perde ve enerji eğrileri, 5. dereceden Legendre polinomlarıyla yakınsandı. Verilen eğriye uyan en iyi Legendre katsayılarını hesaplamak için *numpy* kütüphanesindeki *legfit* fonksiyonu kullanıldı. Bu fonksiyon, verinin Legendre serisine uyumu için en küçük kareler yöntemini kullanmaktadır. Böylece normalize edilmiş gerçek perde ve enerji eğrisiyle, tahmin edilen polinom arasındaki farkı en aza indiren Legendre katsayıları hesaplanmış olur. Yakınsayan polinomun katsayıları öznitelik vektörü olarak kullanıldı.

Böylece her segmentteki perde ve enerji eğrisi için 6'şar katsayı elde edildi. Her segmentin süresi, çerçeve sayısı cinsinden bulunarak toplamda 13 katsayılı öznitelik vektörü oluşturuldu. Şekil 5.12'deki cümlede 14 segment bulunduğundan 13×14 ebatlı öznitelik matrisi oluşturulmuştur. Her cümlenin segment sayısı farklı olduğundan matrisin boyutu değişkenlik göstermektedir.



Şekil 5.12 Segmentlere ayırma ve ünlü kimliklerinin bulunması.

5.4.2.3.2. Ağız Profilleme (Dil Modeli)

Bu çalışmada, her segmentteki ünlü kimliği bulunarak [137] çalışmasındaki ayrık birimler modeli iyileştirilmiştir. İzlenen adımlar aşağıdadır.

5.4.2.3.2.1. Ünlü Kimliğinin Bulunması

SpeechRate ile hece ortaları yani ünlü seslerin yeri bulunmuştu (Şekil 5.12). Ünlüler Praat programı ile etiketlendi. Ünlü sesin bulunduğu çerçevenin etrafındaki $(-5, +5)$ toplam 10 çerçevenin 39 boyutlu MFCC öznitelikleri çıkartıldı. Böylece her ünlü için 39×10 ebatlı öznitelik matrisi elde edildi. MFCC öznitelikleri şöyle çıkartıldı: Konuşma sinyaline hızlı Fourier dönüşümü uygulanarak frekans spektrumu elde edildi ve spektruma 40 kanallı Mel süzgeç bankası uygulandı. 25 ms Hamming penceresi 10 ms örtüşme süresiyle kullanıldı. 13 MFCC katsayısına ek olarak birinci ve ikinci türevlerden gelen 26 katsayı da hesaplanarak toplamda 39 katsayı elde edildi. Daha sonra ünlü seslerin öznitelik matrisleri ve etiketleri kullanılarak ünlü kimliklendirici geleneksel ileri beslemeli sinir ağı eğitildi.

İleri beslemeli sinir ağındaki katmanların düğüm sayıları şöyledir: 390-200-100-50-8. Beş katmanlı sinir ağının gizli katmanlarında sigmoid aktivasyon fonksiyonu, çıkış katmanında ise olasılıkları elde etmek için softmax fonksiyonu kullanıldı. Eğitim çapraz-entropi ölçütüne göre SGD algoritmasıyla yapıldı.

Test aşamasında, aynı şekilde cümleler SpeechRate betiğinden geçirilerek hece ortaları bulundu. Ünlü sesin etrafındaki 10 çerçevenin MFCC katsayıları hesaplandı ve sinir ağının girişine verildi. Böylece sinir ağı, test aşamasında % 92 doğruluk oranı sağlamıştır. Bu oran, ünlü kimliğinin büyük ölçüde doğru tespit edildiğini göstermektedir. Burada her ağız için değil, bütün ağızlar için ortak bir ünlü sınıflayıcı yapılmıştır. Türkçede 8 harfe karşılık 8 ünlü fonem bulunmaktadır. Bu yüzden ileri beslemeli sinir ağının çıkış katmanı 8 sınıflıdır.

5.4.2.3.2.2. Ayrık Birimlerin Elde Edilmesi

Legendre'nin 2. polinom katsayısı bir segmentte bulunan eğrinin eğimini, dolayısıyla artma-azalma özelliğini göstermektedir. Bu özellik ayrık sınıfları bulmak için kullanıldı. Perde ve enerji kontörleri Şekil 5.12'deki gibi hizalandığı için bunların birbirilerine göre durumlarını bulmak kolaydır. Bu çalışmada, her segmentteki ünlü kimliklerinin bulunması önerildiğinden ve ötümsüz (unvoiced) bölgelerde ünlü olmadığından Çizelge 5.7'deki ilk 4 ayrık sınıf kullanılmıştır.

Bir cümledeki segmentlerin ortalama süreleri bulunarak eşik değer olarak belirlendi. Bu eşik değerin altında kalan segmentler kısa (S), üstündekiler ise uzun (L) sınıfı olarak işaretlendi. Tanımsız perdeden (kesikli dikey çizgi) sonra gelen ilk segment, hece ortası belirlenmiş ilk ünlüye atandı. Böylece ağız profillemenin sonunda Şekil 5.12'deki cümle, şu ayrık birimler haline getirilmiştir: A1S A2L U1S A4L İ4S E1S E3S A4S E3S E4L İ1S İ1L İ3L A4L.

5.4.2.4. Sınıflandırma

5.4.2.4.1. LSTM ile Sınıflandırma

Yukarıda elde edilen 13 boyutlu Legendre polinom katsayıları LSTM katmanlı yinelemeli sinir ağında kullanıldı. Her cümle farklı segment sayısına (N) sahip olduğundan öznitelik matrisi $13 \times N$ boyutludur. Farklı sayıdaki segmentten ötürü eğitim verileri, boyu 20 olan mini-yığınlara (mini-batch) ayrıldı. Böylece 3 boyutlu ($20 \times 13 \times N$) bir veri LSTM ağına verilmektedir. Mini-yığın içindeki cümleler segment sayısına göre önce sıralandı ve daha sonra sıfırla doldurma (padding) işlemi yapıldı. Cümlelerin uzunluğa göre sıralanmasının nedeni en uzun cümlenin bulunmasıdır. Kısa cümlelerin uzunluğu, en uzun cümleye sıfırla

doldurma işlemi yapılarak eşitlenir. Bu sayede mini-yığın içindeki uzunlukların eşit olması sağlanmıştır.

LSTM ağının giriş katmanı 13 düğümlüdür. LSTM katmanında 100 gizli düğüm vardır ve dizinin son elemanını çıkışa vermektedir. LSTM katmanından sonra 4 (ağız sayısı) düğümlü tam bağlı çıkış katmanı bulunmaktadır. Çıkış katmanında olasılık dağılımını elde etmek için softmax fonksiyonu ve çapraz-entropi hata değerini minimize etmek için SGD algoritması kullanılmıştır.

Test aşamasında eğitim aşamasıyla aynı şartlar oluşturuldu. Mini-yığın boyu 20 olarak belirlendi, sıralama ve sıfırla doldurma işlemi yapıldı. Veri kümesi, eğitim ve test kümeleri olarak 10-katlamalı çapraz doğrulama yöntemiyle ayrıldı. Böylece tüm veri kümesi 10 parçaya ayrıldı ve bu parçaların 9'u eğitim, 1'i de test için kullanıldı. Bu parçaların eğitim ve test olarak ayrılması işlemi 10 defa arka arkaya yapıldı ve bu 10 denemenin ortalaması alınarak sonuç skoru elde edildi.

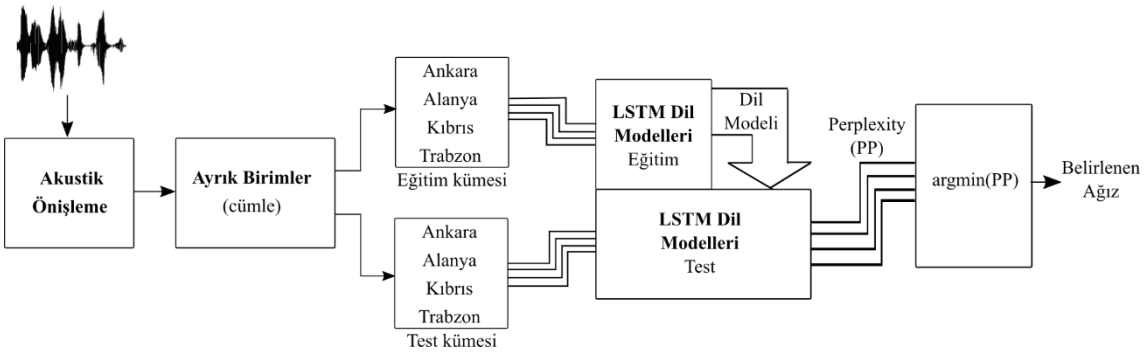
5.4.2.4.2. LSTM ile Dil Modelleme Yapılarak Sınıflandırma

Yukarıda elde edilen ayrık birimler birer sözcük olarak varsayılırsa bu sözcükler ilgili ağız modelini eğitmek için kullanılabilir. Ayrık birimlerin içindeki üç özellik üç ayrı *one-hot* vektörle ifade edilir ve uç uca eklenirse tek bir vektör haline getirilebilir. Bu durumda oluşturulan vektöre *multi-hot* vektör denir. 8 sınıflı ünlü özelliği, 4 sınıflı perde-enerji özelliği ve 2 sınıflı süre özelliği, 14 boyutlu tek bir multi-hot vektörle ifade edilir. Örneğin A1S sözcüğünün vektör gösterimi şöyledir: 00000001|0001|01.

Adami [137] ve Rouas [82], ayrık birimlerin her birini farklı sözcükler olarak n-gram ile modellemişlerdir. Ancak bunlar birbirinden bağımsız farklı sözcükler olarak değerlendirilirse hem vektör boyu uzar, hem de birimler içindeki özelliklerin birbiriyle ilişkisi göz ardı edilir. Örneğin ünlü sesin kimliği ile süresi arasında bir bağlantı söz konusu ise *one-hot* vektör gösterimi ve n-gram ile bu ilişki modellenemeyecektir. Hem birimler içindeki özelliklerin, hem de birimlerin birbirileriyle ilişkisini modelleyebilmek için *one-hot* vektör yerine *multi-hot* vektör kullanılabilir.

LSTM katmanlı sinir ağının giriş ve çıkış katmanları aynı boyuttadır. Sinir ağının girişine ve çıkışına 14 boyutlu *multi-hot* vektör verilmektedir. Örneğin yukarıda elde edilen ayrık

birim cümlesi sinir ağına şöyle verilir: $T = 1$ zamanında girişe *A1S* vektörü verilirken çıkış etiketini *A2L* oluştururur. Aynı şekilde $T = 2$ zamanında girişe *A2L* verilirken çıkış etiketi *U1S* olmaktadır. Bu şekilde her bir zaman adımında bir kelime işlenerek dizinin tamamı işleninceye kadar süreç devam eder (Kısım 5.3'teki fonotaktik modele benzer şekilde). Bu işlem bir ağızdaki bütün eğitim verisi üzerinde tekrar edilmektedir. Bu tekrar ağın tahminleri, eğitim verileriyle tutarlı hale gelene kadar devam etmektedir. Böylece o zamana kadar olan kelimelere göre sonraki kelimenin tahmin edilmesi modellenerek her bir ağız için LSTM dil modeli eğitilir (Şekil 5.13).



Şekil 5.13 Ağızların LSTM dil modeli ile sınıflandırma mimarisi.

LSTM ağının eğitimi ikili çapraz-entropi (Eş. 5.11) ölçütüne göre SGD algoritmasıyla yapıldı ve türevler, zaman boyunca hatanın geri yayılımı (BPTT) algoritmasıyla hesaplandı. BPTT algoritması yinelemeli sinir ağını, verilen zaman adımı sayısı kadar katmanı olan ileri beslemeli sinir ağına çevirerek türevleri hesaplar.

$$CE = -\frac{1}{n}\sum_x \big(y \ln y + (1-y) \ln (1-y)\big) \tag{5.11}$$

Burada n örnek cümle sayısıdır. *Multi-hot* vektör yapısından dolayı çıkış katmanında sigmoid aktivasyon fonksiyonu kullanılmıştır. Başlangıçta ağırlık matrisleri sıfıra yakın değerlerle ilklendirilerek öğrenme hızı $\propto = 0{,}1$ olarak belirlendi. Her 10 örnekten sonra ağın öğrenme kabiliyeti, doğrulama verileriyle test edildi. Bu şekilde doğrulama verisinin olasılık değerlerinin artması durumunda eğitim devam etmekte, aksi halde $\propto$ değeri yarıya düşürülmektedir. Olasılık değerinin önemli oranda artmaması halinde ise eğitim sonlandırılmaktadır. Yinelemeli olarak bağlanan iki gizli katman ve her katmanda 50 LSTM birimi vardır. Zaman adımı $T = 10$ olarak alındı. Bu değer, açık hale getirilen sinir

ağının katman sayısına denk gelmektedir ve hafızada tutulacak sözcük sayısını gösterir. Eğitim, test ve doğrulama için konuşma örnekleri yüzde olarak 80, 10, 10 oranında ayrıldı.

Verilen bir ayrık birim cümlesi için her bir LSTM modelinin *perplexity* metriği Eş. 5.12'deki gibi hesaplanmıştır. CE test verisinin entropisidir.

$$PP = 2^{CE} \tag{5.12}$$

Test için ayrılan ayrık birimler, eğitilen LSTM dil modelinden geçirilerek PP değerleri üretilmektedir. Burada en düşük PP değeri veren dil modeli seçilerek ayrık birimin ait olduğu ağız sınıfı belirlenmiştir (Şekil 5.13). LSTM ağı kurulumu ve hesaplamaları Keras kütüphanesi yardımıyla yapılmıştır.

5.4.3. Bulgular ve Tartışmalar

Kullanılan yöntemlerin belli sürelerdeki test örneklerine göre ürettiği doğruluk oranları Çizelge 5.8'de verilmiştir.

İkinci dereceden Legendre özniteliklerinin beşinci dereceden Legendre özniteliklerine göre daha düşük doğruluk oranı verdiği görülmektedir. Bu durum, Legendre polinom derecelerinin artmasıyla eğriyi temsil gücünün de artmasının bir sonucudur. Katsayılar arttıkça fonksiyon da eğriyi daha iyi yakınsamaktadır.

Ayrık birimlerin LSTM dil modeli ile gerçeklendiği yöntemlerin başarımları, 0,5 s ve 1 s sürelerinde Legendre katsayılarının kullanıldığı yöntemlere göre daha düşüktür. Bu, ayrık birimler için 0,5 s ve 1 s sürelerinin ayırıcı bilgi sağlama açısından yeterli olmadığı şeklinde yorumlanabilir. Ancak uzun test sürelerinde (3 s) bu durum tersine dönmekte ve başarım oranı artmaktadır. Dil modeli içeren yöntemler, n-gram ya da sinir ağı ile yapılmış dil modeli olabilir, tutarlı sonuç üretmesi için daha uzun (> 3) test sürelerine ihtiyaç duymaktadır. Ayrıca 3. yöntemdeki klasik prozodik özelliklere, ünlü kimliğinin eklenmesiyle (4) başarım oranı % 76,2 olmaktadır. Ünlü kimliğinin bulunarak [57], [137] ve [82] sonuçlarının geliştirilmesi bu çalışmanın yeniliklerinden biridir. Bulunan ünlü kimlikleriyle örneğin süre arasındaki örüntü burada önerilen yöntemle modellenebilmektedir.

Çizelge 5.8 Yöntemlerin ürettiği doğruluk oranları (%).

	Yöntemler	Doğruluk oranı (%) 0,5 s	1 s	3 s
1	Legendre (3 katsayı) + LSTM	68,6	69,3	72,5
2	Legendre (6 katsayı) + LSTM	71,5	71,7	**74,9**
3	Perde, enerji, süre + LSTM Dil Modeli	68,1	70,5	75,0
4	Perde, enerji, süre + ünlü + LSTM DM	68,5	71,0	**76,2**
5	Perde, enerji, süre + ünlü + 3-gram DM	65,2	66,5	70,4
6	2 ve 4 birleştirilmesi	69,1	72,6	**78,7**

Prozodik özelliklerin LSTM sinir ağları kullanılarak elde edilen dil modelleriyle gerçeklenmesi ve bunun sonucuna göre sınıflandırma yapılması çalışmanın bir diğer yeniliğidir. LSTM dil modellerinin klasik dil modellerine üstünlüğünü görmek açısından (4)'te kullanılan özellikler 3-gram dil modelleriyle modellenmiştir (5). N-gram dil modeli için 5.3 Kısmında tanıtılan araç [134] kullanılmıştır. Elde edilen sonuçlar sadece LSTM dil modelinin değil diğer modellerin de gerisinde kalmıştır.

En iyi yöntemlerin ürettiği olasılık değerlerinin çarpılmasıyla elde edilen en yüksek çarpım değerini veren sınıfın seçilmesine iki yöntemin birleştirilmesi (fusion) denilmektedir. Dil ve konuşmacı tanıma çalışmalarında sıklıkla bu yola başvurulmaktadır. Bu yaklaşımda, kullanılan yöntemlerin her sınıf için ürettiği olasılıklar eleman elemana çarpılır ve en yükseği veren sınıf seçilir. Altıncı yöntemde (2) ve (4)'ün sonuçları birleştirildiğinde tanıma oranı % 78,7 ile en yüksek orana çıkmaktadır.

Aşağıda ikinci ve dördüncü yöntemlerin 3 s test süresi için karışıklık matrisleri sırasıyla Çizelge 5.9 ve 5.10'da verilmiştir.

Her iki yöntem de Trabzon ağzını diğerlerinden daha iyi ayırırken en düşük oranlar Ankara ağzı için üretilmiştir. Bu sonuç Ankara ağzının, diğer ağızlara göre daha az ayırt edilebilir karakteristiklere sahip olduğunu gösterir. Aynı şekilde, Kıbrıs ve Alanya ağızlarının birbiri arasındaki karışıklık oranlarının daha yüksek olması nedeniyle prozodik olarak birbirine diğerlerinden daha çok benzediği söylenebilir.

Çizelge 5.9 İkinci (2) yöntemin karışıklık matrisi (%).

	Ankara	Alanya	Kıbrıs	Trabzon
Ankara	**73,2**	11,3	8,5	7,0
Alanya	9,5	**75,1**	10,2	5,2
Kıbrıs	8,0	9,6	**74,7**	7,7
Trabzon	8,1	7,5	7,8	**76,6**

Çizelge 5.10 Dördüncü (4) yöntemin karışıklık matrisi (%).

	Ankara	Alanya	Kıbrıs	Trabzon
Ankara	**74,4**	9,2	8,8	7,6
Alanya	8,2	**75,6**	10,7	5,5
Kıbrıs	7,5	9,1	**76,1**	7,3
Trabzon	7,8	7,1	6,4	**78,7**

5.4.4. Prozodik Açıdan Sonuçlar

Prozodik bilginin heceler üzerinde taşındığı bilinmektedir. Segmentleme için temel birim hecedir. Heceler ise içinde bir ünlü ses barındıran birimler olarak kabul edilir. Bundan dolayı bu çalışmada, hece içindeki ünlü kimliğinin bulunması önerilmektedir.

Ayrık birimler içindeki sınıfların kendi aralarındaki ilişki *multi-hot* vektör yapılarıyla gösterilmiştir. Ayrıca ayrık birimler arasındaki uzun dönemli bağımlılıklar n-gramlar kullanılarak gösterilemezler. Burada, uzun dönemli bağımlılığı göstermek için LSTM yinelemeli sinir ağı kullanılmıştır. LSTM sinir ağının, ağızların prozodik profilini elde etmek amacıyla kullanılması ve bu bilgiyle sınıflandırma yapılması bir ilktir. Buna ek olarak LSTM sinir ağı daha önce sınıflandırıcı olarak ham zamansal verilerle denenmiştir, ancak bu tez çalışmasında giriş verisi olarak ham zamansal verilere en uygun polinom öznitelikleri kullanılmıştır.

6. TARTIŞMA VE SONUÇLAR

Bu tez çalışmasında akustik, fonotaktik ve prozodik öznitelikleri bakımından Türkçenin ağızlarının, bilgisayar destekli olarak birbirinden ayırt edilmesi konu edilmiştir. Tasarlanan sistemde denemeler için Türkçenin Ankara, Alanya, Kıbrıs ve Trabzon ağızlarına yönelik bir derlem oluşturulmuştur.

Veri kümesi oluşturma süreci zor ve uzundur. Henüz Türkçenin ağızlarının otomatik olarak sınıflandırılmasına hizmet edebilecek nitelikte bir veri kümesi bulunmamaktadır. Ağız Bilimi çalışanlarca toplanan konuşma kayıtları amaca uygun derlenmiştir. Bu kayıtlar tasnif edilmiş ve üzerinde işlem yapılabilir şekle getirilmiştir. Çalışma sonucunda 2,7 saatlik bir veri kümesi ortaya çıkartılmış ve adına Türkçenin Ağızları Veri Kümesi denilmiştir. Tezdeki deneyler bu veri kümesi üzerinde yapılmıştır.

Diller/Ağızlar arasında ses, ses dizimi, tonlama, vurgu gibi çeşitli seviyelerde dil bilimsel özelliklerden kaynaklanan farklar bulunur. Bu farklar Türkçenin ağızlarında da gözlenmektedir. Bu çalışmada, Türkçenin ağızlarındaki bu farklara dayalı olarak otomatik sınıflandırma işlemleri araştırılmıştır. Sınıflandırma işleminde derin öğrenme teknikleri ele alınmıştır.

Akustik özellikler, konuşma sinyalinin spektral düzeydeki görünümü ile ilgili bilgi vermektedir. Sinyalin özet bilgisi MFCC ve süzgeç bankaları yöntemleriyle çıkartılmıştır. Bu bilgi daha sonra ileri beslemeli ve konvolüsyonel sinir ağlarının girdisini oluşturmak için kullanılmıştır. Yapılan deneyler, Türkçenin ağızlarının akustik olarak sinir ağlarıyla sınıflandırılabileceğini ortaya koymuştur. CNN ile yapılan deneyde 3 saniye süreli konuşma örneklerinin süzgeç bankaları özniteliği kullanılarak % 84,4 doğruluk oranıyla sınıflandırıldığı görülmüştür.

Oluşturulan Türkçenin ağızları derlemi, cümle düzeyinde 2-3 saniyelik birimlere ayrılmıştır. Bu birimler üzerinde yapılan incelemelerde, ağza özgü bilgilerin cümle sonundaki sözcükler üzerinde yoğunlaştığı görülmüştür. Bu sözcüklerin tonlama ve vurgu örgüleri ağızdan ağza değişmektedir. Bu örgüden faydalanmak için cümle sonundaki 0,5 ve 1 saniye uzunluğundaki kısımların öznitelikleri kullanılmıştır. 1 saniyelik test örneğinin CNN ağında kullanılmasıyla % 84,2 doğruluk oranı yakalanmıştır. Bu sonuca göre,

konuşma örneklerinin sadece çok kısa süreli cümle sonlarından Türkçenin ağızlarının sınıflandırılması önemlidir. Bu şekilde bilgisayarın hesap gücünden ve zamandan tasarruf sağlanabilmektedir.

Türkçenin ağızlar derlemi metinden bağımsız rastgele konuşmalardan oluşmaktadır. Bu özelliği, veri kümesindeki konuşmaları monoton yapıdan uzaklaştırmakta ve konuşmacıların ağza özgü detayları daha çok kullanmasını sağlamaktadır. Konuşma tanıma uygulamalarında çokça kullanılan TIMIT veri kümesi ise metne bağımlıdır ve ağızların daha çok fonetik farklılıktan kaynaklanan spektral özelliklerine dayalıdır. Bu iki veri kümesi de cümle düzeyinde 2-3 saniyelik birimlere ayrılmış durumdadır. Türkçe için önerilen cümle sonundaki sabit kısımların sınıflandırmada kullanılması fikri TIMIT için de denenmiştir. Ancak ileri beslemeli ve konvolüsyonel sinir ağlarının Türkçe üzerinde TIMIT'e göre sırasıyla % 2,1 – 2,6 daha iyi başarım oranı sağladığı görülmüştür. Bu farkın nedeni yukarıda açıklanan iki veri kümesi arasındaki farktan kaynaklandığı düşünülmektedir.

Bir kişinin konuşmasında iki bilgi söz konusudur: 1) fiziksel ses yolu özellikleri, 2) dil/ağız özellikleri. Akustik incelemede bu iki bilgi birbirinden ayrılamayabilir. Diğer bir deyişle, akustik farklılığın hangi bilgi türünden kaynaklandığı bilinemez. Ancak bunlar bir kişi için birbirini tamamlayıcı bilgilerdir. Yani bir kişi tespit edilebilirse onun kimliğinden ağzı bölgesi bulunabilir. Bu bakımdan ağız tanıma bağlamında, akustik farklılığın hangi bilgi türünden kaynaklandığının bilinmesine o kadar gereksinim yoktur, sonucuna varılabilir.

İkinci aşamada ağızların fonotaktik özellikleri üzerinde durulmuştur. Bütün sesler her ağızda olabilir ancak bunların bulunma sıklıkları ve birbirini izleyen şekilde dizilmeleri ağızdan ağza değişmektedir. Bu tanımdan hareket edilirse, fonotaktik özelliğin ortaya konulması için konuşulan fonemlerin belirlenmesi ve bunların fonolojik örüntüsünün çıkartılması gerektiği görülmektedir. Fonotaktik için çok popüler olan Paralel PRLM tekniği bu çalışmada kullanılmıştır. Fonemler belirlendikten sonra fonolojik örüntünün (dil modeli) elde edilmesinde N-gram modeller kullanılmaktadır. Bu çalışmada N-gram modeller yerine LSTM katmanlı yinelemeli sinir ağlarıyla dil modeli oluşturma yoluna gidilmiştir. Yapılan deneyler sonucunda LSTM ile oluşturulan dil modelinin N-gram modeline göre yaklaşık % 6 daha iyi doğruluk oranı ürettiği görülmüştür. LSTM'in fonotaktik anlamda dil/ağız tanımada kullanılması bu tez çalışmasının literatüre

katkılarından birisidir. Dil tanıma alanındaki üç çalışma grubunun 2016 tarihli ortak bildirisine [135] göre, fonotaktik için kullanılan N-gram modellemenin yerini ileriki dönemlerde LSTM ağlarının alabileceği öngörülmektedir. Bu tez çalışmasının literatüre katkısı bu bakımdan değerlendirilebilir.

Fonotaktik modelleme ve buna göre sınıflandırma işlemi için uçtan uca bir sistem kurulmamıştır. Bu bakımdan bu sistem üç ayrı işlevin birleşimi olarak düşünülebilir: 1) iki fonem tanıyıcı tarafından fonemlerin elde edilmesi, 2) bu fonemlere dayalı olarak 8 adet birbirinden bağımsız LSTM dil modelinin kurulması ve 3) dil modellerinden hesaplanan skorların ayrı bir sinir ağı ile sınıflandırılması.

Dil bilimsel spektrumun üçüncü katmanında prozodik öznitelikler yer almaktadır. Prozodik özniteliklere suprasegmental veya parçalar üstü öznitelikler de denilmektedir. Tonlama, vurgu ve ritim dilin prozodisini oluşturmaktadır. Üst düzeydeki bu bilgilerin sinyal düzeyindeki karşılıkları temel frekans (f_0), enerji ve sürenin birleşimidir. Bu tez çalışmasında alt düzey ölçümlerle Türkçenin ağızlarının prozodik örüntüsü çıkartılarak sınıflandırma yolu benimsenmiştir. Bunun için ilk olarak bu ölçüm değerlerinin özet bilgisi çıkartıldı. Eldeki veri noktalarının yakınsamasını sağlamak için Legendre polinom katsayıları kullanıldı. Elde edilen katsayılar hece düzeyinden hesaplandığı için bütün bir cümle zamansal olarak LSTM sinir ağına verilebilir. Bu şekilde her zaman adımında 13 Legendre katsayısının kullanılarak LSTM ağı ile çoka-bir tarzda sınıflandırma sonucunda 3 saniyelik test cümleleri için % 74,9 doğruluk oranı yakalanmıştır. Her bir zaman adımında kullanılan Legendre katsayısının artması durumunda sınıflandırma başarımının da arttığı görülmüştür.

Diğer bir uygulamada, ağızlardaki her cümlenin prozodik örüntüsünün çıkarılması için Adami'nin [137] ayrık birimler yaklaşımına göre hareket edilmiştir. Hece ortalarındaki (syllable nuclei) ünlü kimliklerinin ayrık birimlere eklenmesi önerilmektedir. Ağızlardaki her cümlenin ayrık birimlere dönüştürülmesi ve artık ayrık birim olarak modellenebilen ağızların profilinin çıkartılması düşünülebilir. Ayrık birimler kullanılarak LSTM ile, tıpkı fonotaktik kısmında anlatıldığı gibi, her ağzın dil modelinin çıkartılması ve bunun sonucunda hesaplanan skora göre ağızların sınıflandırılması bu çalışmanın yeniliklerinden biridir. Bu şekilde, ünlü eklenmiş ayrık birimler ve LSTM dil modelleri sonucunda % 76,2 sınıflandırma oranı bulunmuştur. Ünlü kimliğinin eklenmesinin % 1,2 iyileştirme sağladığı

görülmüştür. Ayrıca önerilen modelin N-gramlı modelden yaklaşık % 6 daha iyi olduğu gözlenmiştir.
Daha önce yapılan çalışmalarda [57, 82, 137] heceler geniş anlamda ünlü ve ünsüz içermesi bakımından modellenmiştir. Ağızlardaki ortak ünlü kimliklerinin bulunarak modellemeye gidilmesi bu çalışmaya özgüdür.

Yöntemlerden elde edilen skorların birleştirilmesi (fusion) dil/ağız/konuşmacı tanıma uygulamalarında sıklıkla görülmektedir. Bu çalışmada, yukarıda bahsedilen iki yöntemin bir ağız test örneği için ürettiği olasılıklar çarpılarak en yüksek oranı veren ağız seçilmiştir. Bu şekilde birleştirme yoluyla % 2,5 daha iyi bir doğruluk oranı elde edilmiştir.

Aşağıdaki çizelgede akustik, fonotaktik ve prozodik açıdan elde edilen en iyi sonuçlar birlikte verilmiştir.

Çizelge 6.1 Farklı öznitelik seviyelerinden elde edilen, süreye bağlı sonuçlar

Öznitelik seviyesi	0,5 s	1 s	3 s
Akustik (LogMel-CNN)	83,7	84,0	84,4
Fonotaktik (PPRLM-LSTM)	-	84,2	85,1
Prozodik (Perde, enerji, süre + ünlü + LSTM DM)	68,5	71,0	76,2

Çizelge 6.1'e göre, Türkçenin ağızlarının sınıflandırılmasına yönelik yapılan uygulamalar sonucunda üretilen en iyi doğruluk oranlarının sırasıyla fonotaktik, akustik ve prozodik yöntemlerden elde edildiği söylenebilir. Buradan hareketle fonotaktik yöntemin en ayırt edici bilgileri sağladığı düşünülmektedir. Akustik özniteliklerin, en düşük seviyede olmasına ve en kolay elde edilmesine rağmen, yine de iyi sonuçlar verdiği görülmektedir. Prozodi genelde diller ve hatta dil grupları arasındaki ayrımlar söz konusu olduğunda daha etkiliyken ayrımın az olduğu ağızlarda daha düşük başarım oranlarına neden olmaktadır. Ancak yine de Türkçenin ağızlarının prozodi ile sınıflandırılabileceği bu çalışmada gösterilmiştir.

Fonotaktik öznitelikler için 0,5 s süreli konuşma örneklerinden ayırt edici derecede bilgi elde edilemediği için bu sürede denemeler yapılmamıştır.

Ağızlardan elde edilen üç öznitelik seviyesinin birlikte kullanıldığı (fusion) büyük bir sistem tasarlanabilir. Böyle bir sistem genelde iki yolla gerçeklenmektedir. Ya bütün seviyelerde bulunan öznitelikler birleştirilerek tek bir sınıflandırıcıyla ağızların ayırt edilmesi sağlanır, ya da her bir öznitelik seviyesi için kullanılan ayrı sınıflandırıcıların sonuçlarının birleştirilmesi yoluna gidilir. Bu tez çalışmasında konuşma örneklerinden elde edilen özniteliklerin türleri ve formatları farklı olduğundan ikinci yol daha uygun olabilir. Böyle büyük bir sistemin ağız sınıfları arasındaki farklılıkları daha iyi yakalayabileceği ve tanıma performansını arttıracağı düşünülmektedir. Prozodik açıdan uygulanan iki modelin sonuçları birleştirildiğinde her iki modelden daha iyi sonuçlar elde edildiği görülmüştür.

İleride Yapılması Planlanan Çalışmalar

Ağızlara ilişkin veri kümesinin, Türkçenin diğer ağızlarını da içerecek şekle getirilmesi planlanmaktadır. En azından Türkiye'nin bütün bölgelerini ya da etki alanı geniş ağızlarını kapsaması sağlanabilir. Bu bakımdan eldeki verilerin de Internet ortamına aktarılarak başka araştırmacıların katkı yapmasına yönelik bir alt yapı kurulması düşünülmektedir. Böylece hem bu alanda önemli bir eksiklik giderilebilir hem de dil bilimci ve bilgisayar bilimcilerin çalışmalarına kaynaklık sağlanabilir.

Veri kümesinin yukarıda anlatıldığı biçimiyle genişletilmesinin yanı sıra veri arttırma (data augmentation) yöntemleriyle de eldeki veriler çoğaltılabilir. Veri arttırma yoluyla artan verinin ağız tanıma performansını nasıl etkileyeceği incelenmelidir. Bu yöntemin özellikle görüntü işlemede performans artışına neden olduğu bilinmektedir.

Öğrenmenin transfer edilmesi (transfer learning), genelde küçük veri kümelerinin olduğu durumlarda uygulanan bir yöntemdir. Daha büyük veri kümeleri üzerinde sinir ağları eğitilir. Sonrasında, öğrenilen parametreler sabit tutularak bu sefer küçük veri kümesiyle işlemler yapılır. Sinir ağının ilk birkaç katmanının, problemin genel özelliklerini öğrendikleri bilinmektedir. Bu yüzden ilk birkaç katmanın parametreleri diğer veri kümesi için kurulan sinir ağının parametrelerini oluşturmaktadır ve asıl öğrenme sonraki katmanlarda olmaktadır. Böylece Türkçe ağız tanıma için bu yöntemin sonuçları araştırılabilir.

Akustik modellemede özellikle otomatik kodlayıcı sinir ağları kullanılarak özniteliklerin çıkartılması yoluna gidilmesi planlanmaktadır. Bu şekilde tıkanıklık katmanlarında elde edilene benzer sonuçlar gözlenebilir.

Fonotaktik modelleme için kullanılan yöntemin ilk aşamasında fonem tanıyıcılar yer almaktadır. Türkçenin standart ağzı (İstanbul) için bir fonem tanıyıcının geliştirilmesi gereklidir. Bunun diğer fonem tanıyıcılarla birlikte kullanılması durumunda, Paralel PRLM yönteminde elde edilen performansın arttırılabileceği öngörülmektedir. Ayrıca Türkçenin diğer ağızlarındaki karakteristik seslerin de eklenerek genel anlamda Türkçe ağızları üzerinde çalışan bir genel fonem tanıyıcı yapılabilir.

Çıkartılan prozodik modellerle eğitilen LSTM sinir ağlarının ağızlara özgü prozodik bilgiyi üretmesi gelecekte yapılması planlanan çalışmalar arasındadır. Böylece Türkçenin ağızlarının karakteristiği, enstrümanlarla çıkartılan seslerle benzetilebilir. Bunun da ötesinde bu sesler notalara dönüştürüldüğünde herhangi bir enstrümanla örneğin Trabzon ağzının melodisi oluşturulabilir.

KAYNAKLAR

[1] P. Carr, "Phonology." The Macmillan Press, p. 340, **1993**.

[2] H. Dawson and M. Phelan, *Language Files: Materials for an Introduction to Linguistics (12th ed.)*. The Ohio State University Press, **2016**.

[3] R. Lass, *Phonology: An Introduction to Basic Concepts*. Cambridge University Press, **1998**.

[4] E. Y. Ceylan, "Ana Türkçede Kapalı e Ünlüsü," *Türk Dil. Araştırmaları*, pp. 151–165, **1991**.

[5] N. Demir and E. Yılmaz, *Türkçe Ses Bilgisi*. Anadolu Üniversitesi Açıköğretim Fakültesi Yayınları, **2011**.

[6] S. Eker, "Türkçenin Sesbirimleri ve Belirgin Altsesbirimleri," *İlmi Araştırmalar*, vol. 24, pp. 23–42, **2007**.

[7] N. Demir, "Ağız Terimi Üzerine," *Türkbilig*, pp. 105–116, **2002**.

[8] N. Demir, "Türkçe Ağız Araştırmalarında Bazı Yöntem Sorunları," *Diyalektoloji Derg.*, no. 4, pp. 1–8, **2012**.

[9] Z. Korkmaz, *Güney-Batı Anadolu Ağızları Ses Bilgisi (Fonetik)*. Ankara, **1956**.

[10] A. Akar, "Ağız Araştırmalarında Yöntem Sorunları," *Türkoloji Derg*., no.2, p. 13, **2006**.

[11] L. Karahan, *Anadolu Ağızlarının Sınıflandırılması*. Türk Dil Kurumu Yayınları, **1996**.

[12] L. Karahan, "Türkiye Türkçesi ağızlarında ñ y değişimi," *Gazi Eğitim Fakültesi Derg.*, pp. 99–105, **1999**.

[13] W. S. McCulloch and W. H. Pitts, "A logical calculus of the ideas immanent in nervous activity," *Bull. Math. Biophys.*, vol. 5, pp. 115–133, **1943**.

[14] F. Rosenblatt, "The Perceptron: A probabilistic model for information storage and organization in the brain," *Psychol. Rev.*, vol. 65, no. 6, pp. 386–408, **1958**.

[15] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning Internal Representations by Error Propagation," *Parallel Distrib. Process. Explor. Microstruct. Cogn.*, vol. 1, pp. 318–362, **1986**.

[16] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A Fast Learning Algorithm for Deep Belief Nets," *Neural Comput.*, vol. 18, pp. 1527–1554, **2006**.

[17] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy Layer-Wise Training of Deep Networks," in *NIPS*, **2007**.

[18] Y. Bengio and Y. LeCun, "Scaling Learning Algorithms towards AI," *Large Scale*

Kernel Mach., **2007**.

[19] D. C. Cireşan, U. Meier, L. M. Gambardella, and J. Schmidhuber, "Deep Big Simple Neural Nets Excel on Hand-written Digit Recognition," *Neural Comput.*, vol. 22, pp. 1–14, **2010**.

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *NIPS*, **2012**.

[21] R. Pascanu, C. Gulcehre, K. Cho, and Y. Bengio, "How to Construct Deep Recurrent Neural Networks," in *ICLR*, **2014**.

[22] G. Xavier and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," *Int. Conf. Artif. Intell. Stat.*, **2010**.

[23] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," *J. Mach. Learn. Res.*, vol. 15, pp. 1929–1958, **2014**.

[24] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *arXiv:1412.6980*, **2014**.

[25] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," *Proc. 27th Int. Conf. Mach. Learn.*, **2010**.

[26] T. N. Sainath, R. J. Weiss, A. W. Senior, K. W. Wilson, and O. Vinyals, "Learning the speech front-end with raw waveform CLDNNs," *Interspeech*, **2015**.

[27] A.-R. Mohamed, G. E. Hinton, and G. Penn, "Understanding How Deep Belief Networks Perform Acoustic Modelling," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, **2012**.

[28] L. Deng and D. Yu, "Deep Learning: Methods and Applications," *Found. Trends Signal Process.*, vol. 7, **2014**.

[29] S. Hochreiter, "Untersuchungen zu dynamischen neuronalen Netzen," *Master's thesis, Inst. für Inform. Tech. Univ. Munchen*, pp. 1–71, **1991**.

[30] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, **1997**.

[31] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, **1998**.

[32] K. Hornik, "Approximation capabilities of multilayer feedforward networks," *Neural Networks*, vol. 4, no. 2, pp. 251–257, Jan. **1991**.

[33] C. Bishop, *Pattern Recognition and Machine Learning*. Springer, **2006**.

[34] L. Bottou, "Large-Scale Machine Learning with Stochastic Gradient Descent," *Proc. COMPSTAT'2010*, pp. 177–186, **2010**.

[35] S. Koturwar and S. Merchant, "Weight Initialization of Deep Neural Networks (DNNs) using Data Statistics," *arXiv:1710.10570v2*, **2018**.

[36] D. C. Cireşan, U. Meier, J. Masci, L. M. Gambardella, and J. Schmidhuber, "Flexible, High Performance Convolutional Neural Networks for Image Classification," *Proc. Twenty-Second Int. Jt. Conf. Artif. Intell.*, **2011**.

[37] O. Abdel-Hamid, A.-R. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional Neural Networks for Speech Recognition," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 22, no. 10, **2014**.

[38] T. Sainath, A. R. Mohamed, B. Kingsbury, and B. Ramabhadran, "Deep convolutional neural networks for LVCSR," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, **2013**, pp. 8614–8618.

[39] G. Montavon, "Deep learning for spoken language identification," *NIPS Work. Deep Learn. Speech Recognit. Relat. Appl.*, pp. 1–4, **2009**.

[40] I. Lopez-moreno, J. Gonzalez-dominguez, O. Plchot, D. Martinez, J. Gonzalez-rodriguez, and P. Moreno, "Automatic language identification using deep neural networks," in *ICASSP*, **2014**, pp. 5337–5341.

[41] A. Satt, S. Rozenberg, and R. Hoory, "Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms," in *Interspeech*, **2017**, pp. 1089–1093.

[42] E. Variani, X. Lei, E. Mcdermott, I. Lopez Moreno, and J. Gonzalez-Dominguez, "Deep Neural Networks For Small Footprint Text-Dependent Speaker Verification," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, **2014**, pp. 4052–4056.

[43] M. Espi, M. Fujimoto, K. Kinoshita, and T. Nakatani, "Exploiting spectro-temporal locality in deep learning based acoustic event detection," *EURASIP J. Audio, Speech, Music Process.*, vol. 2015, p. 26, **2015**.

[44] V. Dumoulin and F. Visin, "A guide to convolution arithmetic for deep learning," *arXiv:1603.07285v2*, **2018**.

[45] "Convolutional Neural Networks." https://developers.google.com/machine-learning/practica/image-classification/convolutional-neural-networks. (Erişim tarihi: **3 Kasım 2018**).

[46] A. Zeyer, P. Doetsch, P. Voigtlaender, R. Schlüter, and H. Ney, "A comprehensive study of deep bidirectional LSTM RNNS for acoustic modeling in speech recognition," in *ICASSP*, **2017**, pp. 2462–2466.

[47] A. Graves and N. Jaitly, "Towards End-To-End Speech Recognition with Recurrent Neural Networks," *JMLR Workshop Conf. Proc.*, vol. 32, no. 1, pp. 1764–1772, **2014**.

[48] I. Sutskever, O. Vinyals, and Q. V Le, "Sequence to sequence learning with neural networks," *Adv. Neural Inf. Process. Syst.*, pp. 3104–3112, **2014**.

[49] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, "On the Properties of Neural Machine Translation: Encoder-Decoder Approaches," **2014**.

[50] T. Mikolov, M. Karafiat, L. Burget, J. Cernocky, and S. Khudanpur, "Recurrent Neural Network based Language Model," *Interspeech*, no. September, pp. 1045–1048, **2010**.

[51] J. L. Elman, "Finding structure in time," *Cogn. Sci.*, vol. 14, no. 2, pp. 179–211, Apr. **1990**.

[52] H. Sak, A. Senior, and F. Beaufays, "Long Short-Term Memory Recurrent Neural Network Architectures for Large Scale Acoustic Modeling," *Interspeech*, no. September, pp. 338–342, **2014**.

[53] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Trans. Neural Networks*, vol. 5, pp. 157–166, **1994**.

[54] I. Sutskever, "Training Recurrent Neural Networks," *PhD Thesis, Univ. Toronto*, pp. 1–101, **2013**.

[55] T. Mikolov, "Statistical Language Models Based on Neural Networks," *PhD Thesis, Brno Univ. Technol.*, pp. 1–133, **2012**.

[56] R. J. Williams and J. Peng, "An efficient gradient-based algorithm for online training of recurrent network trajectories," *Neural Comput.*, vol. 4, pp. 491–501, **1990**.

[57] F. Biadsy, "Automatic Dialect and Accent Recognition and its Application to Speech Recognition," *PhD Thesis,Columbia Univ.*, pp. 1–171, **2011**.

[58] L. Wang, "Automatic Spoken Language Identification," The University of New South Wales, **2009**.

[59] S. F. Chen and J. Goodman, "An Empirical Study of Smoothing Techniques for Language Modeling," *Comput. Speech Lang.*, vol. 13, no. 4, pp. 359–394, **1999**.

[60] Y. Kim, Y. Jernite, D. Sontag, and A. M. Rush, "Character-Aware Neural Language Models," in *Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16) Character-Aware*, **2016**, pp. 2741–2749.

[61] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, "A Neural Probabilistic Language Model," *J. Mach. Learn. Res.*, vol. 3, pp. 1137–1155, **2003**.

[62] T. J. Hazen and V. Zue, "Automatic language identification using a segment-based approach," *3rd Int. Conf. Spok. Lang. Process.*, no. 1, pp. 1307–1310, **1993**.

[63] M. A. Zissman, "Language identification using phoneme recognition and phonotactic language modeling," *IEEE Int. Conf. Acoust. Speech, Signal Process. ICASSP*, no. March 1994, pp. 3503–3506, **1995**.

[64] J. Zhao, H. Shu, L. Zhang, X. Wang, Q. Gong, and P. Li, "Cortical competition during language discrimination," *Neuroimage*, vol. 43, no. 3, pp. 624–633, **2008**.

[65] B. Comrie, *The World's Major Languages*. Oxford Univ. Press, **1990**.

[66] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*,

vol. 35, no. 2. **2008**.

[67] J. Navratil, "Spoken Language Recognition–A step Toward Multilinguality in Speech Processing," *IEEE Trans. Speech Audio Process.*, vol. 9, pp. 678–685, **2001**.

[68] F. Ramus and J. Mehler, "Language identification with suprasegmental cues: a study based on speech resynthesis.," *J. Acoust. Soc. Am.*, vol. 105, no. 1, pp. 512–21, **1999**.

[69] A. Etman and A. A. Louis, "Language and Dialect Identification: A Survey," *IntelliSys 2015 - Proc. 2015 SAI Intell. Syst. Conf.*, pp. 963–970, **2015**.

[70] H. Li, B. Ma, and K. A. Lee, "Spoken language recognition: From fundamentals to practice," *Proc. IEEE*, vol. 101, no. 5, pp. 1136–1159, **2013**.

[71] R. A. Cole, J. W. T. Inouye, Y. K. Muthusamy, and M. Gopalakrishnan, "Language identification with neural networks: A feasibility study," in *IEEE Pacific Rim Conference on Communications, Computers and Signal Processing*, **1989**, pp. 525–529.

[72] S. Nakagawa, Y. Ueda, and T. Seino, "Speaker-independent, text-independent language identification by HMM," in *ICSLP'92*, **1992**, pp. 1011–1014.

[73] A. S. House and E. P. Neuburg, "Toward Automatic Identification of the Languages of an Utterance: Priliminary Methodological Considerations," *J. Acoust. Soc. Am.*, vol. 62, no. 3, pp. 708–713, **1977**.

[74] S. Davis and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Trans. Acoust. Speech Signal Process. Vol ASSP-28, No.4*, no. 4, **1980**.

[75] P. A. Torres-Carrasquillo, E. Singer, M. A. Kohler, R. J. Greene, D. A. Reynolds, and J. R. Deller, "Approaches to Language Identification using Gaussian Mixture Models and Shifted Delta Cepstral Features," in *Int. Conf. Spoken Lang. Process.*, **2002**, pp. 89–92.

[76] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. New Jersey: Prentice Hall, **2008**.

[77] Rong Tong, Bin Ma, Donglai Zhu, Haizhou Li, and Eng Siong Chng, "Integrating Acoustic, Prosodic and Phonotactic Features for Spoken Language Identification," *2006 IEEE Int. Conf. Acoust. Speed Signal Process. Proc.*, vol. 1, p. I-205-I-208, **2006**.

[78] D. Martinez, L. Burget, L. Ferrer, and N. Scheffer, "IVector-based prosodic system for language identification," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, **2012**, pp. 4861–4864.

[79] L. Ferrer, N. Scheffer, and E. Shriberg, "A comparison of approaches for modeling prosodic features in speaker recognition," *2010 IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 4414–4417, **2010**.

[80] C. Y. Lin and H. C. Wang, "Language identification using pitch contour information," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. I, pp. 601–604, **2005**.

[81] A. G. Adami and H. Hermansky, "Segmentation of Speech for Speaker and Language Recognition," in *Eurospeech*, **2003**, pp. 841–844.

[82] J. L. Rouas, "Automatic prosodic variations modeling for language and dialect discrimination," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 15, no. 6, pp. 1904–1911, **2007**.

[83] E. Ambikairajah, H. Li, L. Wang, B. Yin, and V. Sethu, "Language Identification: A Tutorial," *IEEE Circuits Syst. Mag.*, vol. 11, no. 2, pp. 82–108, **2011**.

[84] G. Leonard, "Language Recognition Test and Evaluation," **1980**.

[85] D. Cimarusti and R. B. Ives, "Development of an Automatic Identification System of Spoken Languages: Phase 1," in *ICASSP'92*, **1982**, pp. 1661–1664.

[86] M. Sugiyama, "Automatic Language Recognition Using Acoustic Features," *IEEE Int. Conf. Acoust. Speech, Signal Process.*, pp. 813–816 vol.2, **1991**.

[87] Y. K. Muthusamy, E. Barnard, and R. A. Cole, "Reviewing Automatic Language Identification," *IEEE Signal Process. Mag.*, vol. 11, no. 4, pp. 33–41, **1994**.

[88] Y. Yan, "Development of an Approach to Language Identification Based on Language-dependent Phone Recognition," Oregon Graduate Institue of Science and Technology, **1995**.

[89] E. Wong and S. Sridharan, "Methods to improve Gaussian mixture model based language identification system," in *International Conference on Spoken Language Processing (ICSLP-2002)*, **2002**, pp. 93–96.

[90] A. Hanani, M. J. Russell, and M. J. Carey, "Human and computer recognition of regional accents and ethnic groups from British English speech," *Comput. Speech Lang.*, vol. 27, no. 1, pp. 59–74, **2013**.

[91] K.-P. Li and T. J. Edwards, "Statistical Models for Automatic Language Identification," in *ICASSP'80*, **1980**, pp. 884–887.

[92] T. Schultz, I. Rogina, and A. Waibel, "A. LVCSR-Based Language Identification," in *ICASSP'96*, **1996**, pp. 781–784.

[93] M. A. Zissman, "Comparison of four approaches to automatic language identification of telephone speech," *IEEE Trans. Speech Audio Process.*, vol. 4, no. 1, pp. 31–44, **1996**.

[94] F. Biadsy, J. Hirschberg, and N. Habash, "Spoken Arabic Dialect Identification Using Phonotactic Modeling," *Proc. EACL 2009 Work. Comput. Approaches to Semit. Lang.*, pp. 53–61, **2009**.

[95] H. Soltau, L. Mangu, and F. Biadsy, "From modern standard Arabic to Levantine ASR: Leveraging GALE for dialects," in *ASRU*, **2011**, pp. 266–271.

[96] M. Soufifar, S. Cumani, L. Burget, and J. H. Cernocky, “Discriminative classifiers for phonotactic language recognition with ivectors,” in *ICASSP*, **2012**, pp. 4853–4856.

[97] J. T. Foil, “Language Identification Using Noisy Speech,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, **1986**, pp. 861–864.

[98] F. J. Goodman, A. F. Martin, and R. E. Wohlford, “Improved Automatic Language Identification in Noisy Speech,” in *ICASSP’89*, **1989**, pp. 528–531.

[99] A. Thyme-Gobbel and S. E. Hutchins, “On using prosodic cues in automatic language identification,” *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP ’96*, vol. 3. **1996.**

[100] F. Cummins, F. Gers, and J. Schmidhuber, “Automatic discrimination among languages based on prosody alone,” no. IDSIA-03-99, **1999.**

[101] L. Mary and B. Yegnanarayana, “Extraction and representation of prosodic features for language and speaker recognition,” *Speech Commun.*, vol. 50, no. 10, pp. 782–796, **2008**.

[102] D. Martinez, E. Lleida, A. Ortega, A. Miguel, and Ieee, “Prosodic Features and Formant Modeling for an Ivector-Based Language Recognition System,” *2013 Ieee Int. Conf. Acoust. Speech Signal Process.*, pp. 6847–6851, **2013**.

[103] I. Lopez-Moreno, J. Gonzalez-Dominguez, D. Martinez, O. Plchot, J. Gonzalez-Rodriguez, and P. J. Moreno, “On the use of deep feedforward neural networks for automatic language identification,” *Comput. Speech Lang.*, vol. 40, pp. 46–59, **2016**.

[104] F. Richardson, D. Reynolds, and N. Dehak, “Deep Neural Network Approaches to Speaker and Language Recognition,” *IEEE Signal Process. Lett.*, vol. 22, no. 10, pp. 1671–1675, Oct. **2015**.

[105] D. Martínez, O. Plchot, L. Burget, O. Glembek, and P. Matějka, “Language Recognition in iVectors Space,” in *Interspeech*, **2011**, pp. 861–864.

[106] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, “Front-end factor analysis for speaker verification,” *IEEE Trans. Audio, Speech Lang. Process.*, **2011**.

[107] N. Dehak, P. A. Torres-Carrasquillo, D. Reynolds, and R. Dehak, “Language recognition via Ivectors and dimensionality reduction,” in *Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech*, **2011**, pp. 857–860.

[108] A. Mccree, “Multiclass Discriminative Training of i-vector Language Recognition,” in *IEEE Odyssey: The Speaker and Language Recognition Workshop*, **2014**.

[109] P. Matejka, L. Zhang, T. Ng, H. Mallidi, O. Glembek, J. Ma, and B. Zhang, “Neural Network Bottleneck Features for Language Identification,” in *Odyssey 2014*, **2014**, pp. 299–304.

[110] L. Ferrer, Y. Lei, M. Mclaren, and N. Scheffer, "Study of Senone-Based Deep Neural Network Approaches for Spoken Language Recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 1, pp. 105–116, **2016**.

[111] J. Gonzalez-Dominguez, I. Lopez-Moreno, H. Sak, J. Gonzalez-Rodriguez, and P. J. Moreno, "Automatic Language Identification using Long Short-Term Memory Recurrent Neural Networks," *Interspeech-2014*, pp. 2155–2159, **2014**.

[112] R. Zazo, A. Lozano-diez, J. Gonzalez-dominguez, D. T. Toledano, and J. Gonzalez-rodriguez, "Language Identification in Short Utterances Using Long Short-Term Memory (LSTM) Recurrent Neural Networks," *PLoS One*, **2016**.

[113] R. Li, H. Mallidi, L. Burget, O. Plchot, and N. Dehak, "Exploiting Hidden-Layer Responses of Deep Neural Networks for Language Recognition," in *Interspeech*, **2016**, pp. 3265–3269.

[114] J. Pešán, L. Burget, and J. Cernocky, "Sequence Summarizing Neural Networks for Spoken Language Recognition," in *Interspeech*, **2016**, pp. 3285–3288.

[115] H. Behravan, V. Hautamäki, and T. Kinnunen, "Factors affecting i-vector based foreign accent recognition: A case study in spoken Finnish," *Speech Commun.*, vol. 66, pp. 118–129, **2015**.

[116] Q. Zhang and J. H. L. Hansen, "Dialect recognition based on unsupervised bottleneck features," in *Interspeech*, **2017**, pp. 2576–2580.

[117] J. H. L. Hansen and G. Liu, "Unsupervised accent classification for deep data fusion of accent and language information," *Speech Commun.*, vol. 78, pp. 19–33, **2016**.

[118] A. Ali, N. Dehak, P. Cardinal, S. Khurana, S. H. Yella, J. Glass, P. Bell, and S. Renals, "Automatic dialect detection in Arabic broadcast speech," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, **2016**, vol. 08–12–Sept, pp. 2934–2938.

[119] E. Michon, M. Q. Pham, J. Crego, and J. Senellart, "Neural Network Architectures for Arabic Dialect Identification," in *Proceedings ofthe Fifth Workshop on NLP for Similar Languages, Varieties and Dialects*, **2018**, pp. 128–136.

[120] G. Işık and H. Artuner, "A Dataset For Turkish Dialect Recognition and Classification with Deep Learning," in *26. IEEE Signal Processing and Communications Applications Conference (SIU)*, **2018**.

[121] J. S. Garofolo, L. F. Lamel, W. M. Fischer, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "Acoustic-Phonetic Continuous Speech Corpus," vol. 0, no. January 1993, pp. 1–94, **1993**.

[122] N. Demir, "Ağız Araştırmalarında Kaynak Kişi Meselesi," *Folk. Prof. Dr. Dursun Yıldırım Armağanı*, p. 11, **1998**.

[123] "Sound eXchange software." http://sox.sourceforge.net/. (Erişim tarihi: **14 Şubat 2018**).

[124] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [Computer program]," **2018**. http://www.praat.org/. (Erişim tarihi: 3 Şubat 2018).

[125] K. E. Wong, "Automatic Spoken Language Identification Utilizing Acoustic and Phonetic Speech Information," Queensland University of Technology, **2004**.

[126] A. Lozano-diez, R. Zazo, D. T. Toledano, and J. Gonzalez-Rodriguez, "An analysis of the influence of deep neural network (DNN) topology in bottleneck feature based language recognition," *PLoS One*, vol. 12, no. 8, **2017**.

[127] S. Ganapathy, K. Han, S. Thomas, M. Omar, M. Van Segbroeck, and S. S. Narayanan, "Robust Language Identification Using Convolutional Neural Network Features," in *Interspeech*, **2014**.

[128] P. Ghahremani, V. Manohar, D. Povey, and S. Khudanpur, "Acoustic modelling from the signal domain using CNNs," in *Interspeech*, **2016**.

[129] F. Chollet, "Keras," *Github*, **2015**. https://github.com/fchollet/keras. (Erişim tarihi: 15 Kasım 2017).

[130] B. Mcfee, C. Raffel, D. Liang, D. P. W. Ellis, M. Mcvicar, E. Battenberg, and O. Nieto, "librosa: Audio and Music Signal Analysis in Python," *Proc. 14th Python Sci. Conf.*, no. Scipy, pp. 1–7, **2015**.

[131] M. A. Zissman and E. Singer, "Automatic language identification of telephone speech messages using phoneme recognition and N-gram modeling," in *ICASSP'94*, **1994**.

[132] P. Matějka, P. Schwarz, J. Cernock, and P. Chytil, "Phonotactic Language Identification using High Quality Phoneme Recognition," *Eurospeech2005*, **2005**.

[133] P. Matejka, L. Burget, P. Schwarz, and J. Cernocky, "Brno University of Technology System for NIST 2005 Language Recognition Evaluation," in *IEEE Odyssey: The Speaker and Language Recognition Workshop*, **2006**, pp. 57–64.

[134] L. Zhang, "N-Gram Extraction Tools," **2003**. https://homepages.inf.ed.ac.uk/lzhang10/ngram.html. (Erişim tarihi: 7 Ağustos 2018).

[135] K. A. Lee, H. Li, L. Deng, V. Hautamäki, W. Rao, X. Xiao, A. Larcher, H. Sun, T. H. Nguyen, G. Wang, A. Sizov, J. Chen, I. Kukanov, A. H. Poorjam, T. N. Trong, C.-L. Xu, H.-H. Xu, B. Ma, E.-S. Chng, and S. Meignier, "The 2015 NIST Language Recognition Evaluation: the Shared View of I2R, Fantastic4 and SingaMS," *Interspeech*, pp. 3211–3215, **2016**.

[136] N. Dehak, P. Dumouchel, and P. Kenny, "Modeling prosodic features with joint factor analysis for speaker verification," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 15, no. 7, pp. 2095–2103, **2007**.

[137] A. G. Adami, "Modeling prosodic differences for speaker recognition," *Speech Commun.*, vol. 49, no. 4, pp. 277–291, **2007**.

[138] R. W. M. Ng, T. Lee, and C. Leung, "Spoken Language Recognition With Prosodic

Features," *IEEE Trans. Audio. Speech. Lang. Processing*, vol. 21, no. 9, pp. 1841–1853, **2013**.

[139] R. Fernandez, A. Rendel, B. Ramabhadran, and R. Hoory, "Prosody Contour Prediction with Long Short-Term Memory, Bi-Directional, Deep Recurrent Neural Networks," *Proc. Interspeech*, no. September, pp. 2268–2272, **2014**.

[140] S. Sinha, A. Jain, and S. S. Agrawal, "Acoustic-Phonetic Feature Based Dialect Identification in Hindi Speech," *Int. J. Smart Sens. Intell. Syst.*, vol. 8, no. 1, pp. 235–254, **2015**.

[141] M. Kockmann, L. Burget, and J. H. Cernocky, "Investigations into prosodic syllable contour features for speaker recognition," in *ICASSP 2010*, **2010**, pp. 4418–4421.

[142] L. Ferrer, H. Bratt, C. Richey, H. Franco, V. Abrash, and K. Precoda, "Classification of lexical stress using spectral and prosodic features for computer-assisted language learning systems," *Speech Commun.*, vol. 69, pp. 31–45, **2015**.

[143] N. H. de Jong and T. Wempe, "Praat script to detect syllable nuclei and measure speech rate automatically," *Behav. Res. Methods*, vol. 41, no. 2, pp. 385–390, **2009**.

HACETTEPE ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
~~YÜKSEK LİSANS~~/DOKTORA TEZ ÇALIŞMASI ORJİNALLİK RAPORU

HACETTEPE ÜNİVERSİTESİ
FEN BİLİMLER ENSTİTÜSÜ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI BAŞKANLIĞI'NA

Tarih: 07 / 02 / 2019

Tez Başlığı / Konusu: TÜRKÇE AĞIZLARIN TANINMASINDA DERİN ÖĞRENME TEKNİĞİNİN KULLANILMASI
Yukarıda başlığı/konusu gösterilen tez çalışmamın a) Kapak sayfası, b) Giriş, c) Ana bölümler d) Sonuç kısımlarından oluşan toplam 90 sayfalık kısmına ilişkin, 07/02/2019 tarihinde ~~şahsım~~/tez danışmanım tarafından *Turnitin* adlı intihal tespit programından aşağıda belirtilen filtrelemeler uygulanarak alınmış olan orijinallik raporuna göre, tezimin benzerlik oranı %2'dir.

Uygulanan filtrelemeler:

1- Kaynakça hariç
2- Alıntılar hariç/~~dâhil~~
3- 5 kelimeden daha az örtüşme içeren metin kısımları hariç

Hacettepe Üniversitesi Fen Bilimleri Enstitüsü Tez Çalışması Orjinallik Raporu Alınması ve Kullanılması Uygulama Esasları'nı inceledim ve bu Uygulama Esasları'nda belirtilen azami benzerlik oranlarına göre tez çalışmamın herhangi bir intihal içermediğini; aksinin tespit edileceği muhtemel durumda doğabilecek her türlü hukuki sorumluluğu kabul ettiğimi ve yukarıda vermiş olduğum bilgilerin doğru olduğunu beyan ederim.

Gereğini saygılarımla arz ederim.

07.02.2019
Tarih ve İmza

Adı Soyadı: Gültekin IŞIK
Öğrenci No: H11264943
Anabilim Dalı: Bilgisayar Mühendisliği Anabilim Dalı
Programı: Bütünleşik Doktora
Statüsü: ☐ Y.Lisans ☐ Doktora ☒ Bütünleşik Dr.

DANIŞMAN ONAYI

UYGUNDUR.

Doç. Dr. Harun ARTUNER

ÖZGEÇMİŞ

Kimlik Bilgileri

Adı Soyadı : Gültekin IŞIK

Doğum Yeri : Iğdır

Medeni Hali : Evli

E-posta : tekinler@gmail.com

Adresi : Iğdır Üniversitesi Bilgisayar Mühendisliği Bölümü, Suveren, Iğdır

Eğitim

Lisans : Fırat Üniversitesi Bilgisayar Mühendisliği

Yüksek Lisans : -

Büt. Doktora : Hacettepe Üniversitesi Bilgisayar Mühendisliği

Yabancı Dil ve Düzeyi

İngilizce : KPDS - 72 (2010 - Güz)

İş Deneyimi

Hacettepe Üniversitesi : Araştırma Görevlisi, 2011 - 2018

Iğdır Üniversitesi : Araştırma Görevlisi, 2018 – Devam

Deneyim Alanları

Otomatik Konuşma Tanıma, Dil/Ağız Tanıma, Dil Modelleme, Yapay Öğrenme, Derin Öğrenme, Sinyal İşleme

Tezden Üretilmiş Yayınlar

- G. Işık and H. Artuner, "Turkish dialect recognition in terms of prosodic by long short-term memory neural networks", Journal of the Faculty of Engineering and Architecture of Gazi University, (Accepted).
- G. Işık and H. Artuner, "Turkish dialect recognition using acoustic and phonotactic features in deep learning architectures", Turkish Journal of Electrical Engineering & Computer Sciences, (Submitted).

Tezden Üretilmiş Tebliğ ve/veya Poster Sunumu ile Katıldığı Toplantılar

- G. Işık and H. Artuner, "A Dataset For Turkish Dialect Recognition and Classification with Deep Learning", 26. IEEE Signal Processing and Communications Applications Conference (SIU), May 2018, İzmir.

Ayrıca tezle ilişkili şu çalışma da bulunmaktadır:

- G. Işık and H. Artuner, "Recognition of Radio Signals with Deep Learning Neural Networks", 24. IEEE Signal Processing and Communications Applications Conference (SIU), May 2016, Zonguldak.