MARMARA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

GERÇEK ZAMANLI TÜRKÇE KONUŞMA TANIMA

EYÜP ENSAR KALAYCI

YÜKSEK LİSANS TEZİ
Bilgisayar Mühendisliği
Anabilim Dalı
Bilgisayar Mühendisliği Programı

DANIŞMAN
Dr. Öğr. Üyesi Anıl BAŞ

İSTANBUL, 2023

MARMARA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

GERÇEK ZAMANLI TÜRKÇE KONUŞMA TANIMA

EYÜP ENSAR KALAYCI
523619024

YÜKSEK LİSANS TEZİ
Bilgisayar Mühendisliği
Anabilim Dalı
Bilgisayar Mühendisliği Programı

DANIŞMAN
Dr. Öğr. Üyesi Anıl BAŞ

İSTANBUL, 2023

TEŞEKKÜR

Bu tezi ve sonuçlarını mümkün kılan açık veri kaynak sağlayıcılarına ve üreticilerine, tez kapsamında hiçbir desteğini esirgemeyen kıymetli hocam Dr. Öğr. Üyesi Anıl BAŞ'a teşekkür ederim. Çalışmamı bana maddi ve manevi her açıdan yardımcı olan sevgili anne ve babama ithaf ediyorum.

Temmuz 2023 **Eyüp Ensar KALAYCI**

İÇİNDEKİLER

ÖZET

Konuşma tanıma, konuşulan dilin bilgisayar tarafından tanınmasını ve metne çevrilmesini sağlayan teknolojiler geliştiren bilgisayar bilimi ve hesaplamalı dilbilimin disiplinler arası bir alt alanıdır. Son 30 yıl içerisinde büyük ölçüde gelişmiş ve kullanımını etkili şekilde artırmış teknolojiler arasındadır. Günümüzde bu teknolojiden sayısız alanda destek alınmaktadır; araç-içi sistemler, tıp, raporlama, askeri alanlarda özellikle hava araçların, telefon ve uygulamalarından olan interaktif sesli yanıt sistemleri, ev otomasyon sistemleri ayrıca engele sahip insanların hayatlarını kolaylaştırma uygulama alanlarıdır. Literatürde Otomatik ASR veya STT olarak kısaltılan konuşma tanıma teknolojisi, Türkçe için de çözüm ve iyileştirme bekleyen açık problemler arasında görülebilir.

2006 ve sonrası makinelerin hızlanması ve Sinir Ağları için eğitim sorunlarının çözümü ile bu alanda Sinir Ağları son teknoloji konuma geldi ve sonrasında özellikle konuşma tanıma gibi anlık bilginin önceki girişlere bağlı olduğu problemlerin çözümü için icat edilen Yinelemeli Sinir Ağları tercih edildi. Günümüzde ise uçtan-uca olarak isimlendirilen; kompleks farklı modellerin bir arada kullanılmasının aksine yalnızca bir tek model ile konuşma tanımaya çözüm arayan modeller tercih edilmekte ve bu yöntem geçerli son teknolojiye ev sahipliği yapmaktadır.

Bu çalışmada ise Türkçe için gerçek zamanlı konuşma tanımanın ele alınması ve yüksek performansla çalışan son teknoloji örneğin sunulması üzerinde araştırma yapılmış ve uçtan-uca yöntemler tercih edilmiştir. Bu kapsamda eğitimler ve testler için kullanılan veri seti sıfırdan derlenmiş ve veri artırım yöntemleri kullanılmadan 6000 saatten fazla Türkçe konuşma ses veri derlenmiştir.

Çalışma kapsamında geleneksel modeller ve uçtan-uca modeller eğitilmiş, performans farkları ortak bir veri seti üzerinden sunulmuştur. Bu noktada geleneksel yöntemlerde Türkçenin sondan eklemeli bir oluşu sebebiyle sıklıkla kendini gösteren sözlük dışı kalma problemi incelenmiş, yeni yöntemlerin bu konudaki performansları araştırılmıştır. Uçtan-uca konuşma tanıma modelleri olarak göze çarpan Transformer ve devamında geliştirilen Conformer tezde ana konu olarak ele alınarak ve konuşma tanıma için gerekli olan ses aktivitesi dedektörü, gürültü azaltma veya bastırma gibi konular üzerinde de araştırmalar yapılmıştır.

ABSTRACT

Speech recognition is an interdisciplinary subfield of computer science and computational linguistics that focuses on developing technologies for recognizing and translating the spoken language into text by computers. It has greatly advanced over the past 30 years and is among the technologies that have significantly increased their usage. Nowadays, this technology is utilized in numerous fields, including in-vehicle systems, medicine, reporting, military applications, especially in the context of aerial vehicles, interactive voice response systems in phones and applications, home automation systems, and applications that facilitate the lives of individuals with disabilities. In the literature, Automatic Speech Recognition (ASR) or Speech-to-Text (STT) technology, abbreviated as ASR or STT, can be seen as one of the open problems that require solutions and improvements for Turkish.

Since 2006, with the increase in computational power and the practical resolution of issues related to Neural Networks, Neural Networks have become the state-of-the-art technology in this field. Subsequently, Recurrent Neural Networks (RNNs) were invented, particularly for solving problems where real-time information depends on previous inputs, such as speech recognition. Nowadays, end-to-end models, which are referred to as "end-to-end," are preferred. Unlike using complex different models together, these models seek solutions for speech recognition using only a single model, and this approach hosts the current state-of-the-art technology.

In this study, real-time speech recognition for Turkish is addressed, and research is conducted on presenting state-of-the-art technology that performs with high efficiency. End-to-end methods are preferred in this context. Within this scope, a data set used for training and testing is compiled from scratch, and more than 6,000 hours of Turkish speech data is collected without using data augmentation methods.

Traditional models and end-to-end models are trained and their performance differences are presented on a common data set. In this regard, the out-of-vocabulary issue that frequently arises due to the agglutinative nature of Turkish in traditional methods is examined, and the performance of new methods in this regard is investigated. Transformer, which stands out as an end-to-end speech recognition model, and its subsequent development, Conformer, are considered as the main topics in this thesis. Additionally, research is conducted on topics related to speech recognition, such as speech activity detection, noise reduction or suppression.

SEMBOLLER

GB **:** Gigabyte

Hz **:** Hertz

MB **:** Megabyte

TB **:** Terabyte

kHz **:** Kilo Hertz

ms **:** Milisaniye

s **:** Saniye

KISALTMALAR

AAC : Advanced Audio Coding

ANN : Artificial Neural Networks

CTC : Connectionist Temporal Classification

DARPA : Defence Advanced Research Projects Agency

DCT : Discrete Cosine Transform

DFT : Discrete Fourier Transform

DNN : Deep Neural Networks

DTW : Dynamic Time Warping

FFT : Fast Fourier Transform

GMM : Gaussian Mixed Model

GRU : Gated Recurrent Unit

HMM : Hidden Markov Model

IFT : Inverse Fourier Transform

LDC : Linguistic Data Consortium

LPC : Linear Predictive Codes

LSTM : Long-Short Term Memory

LVCSR : Large Vocabulary Continious Speech Recognition

M4A : MPEG-4 Audio

MFC : Mel Frequency Cepstrum

MFCC : Mel Frequency Cepstral Coefficient

MP3 : MPEG Audio Layer III

MPEG : Em-peg

OCR : Optical Character Recognition

OOV : Out of Vocabulary

OpenSLR : Open Speech and Language Resources

PCM : Pulse Code Modulation

PER : Phoneme Error Rate

PESQ : Perceptual Estimation of Speech Quality

PLP : Perceptual Linear Prediction

RNN : Recurrent Neural Networks

S16LE : Signed 16 Little Endian

SAD **:** Speech Activity Detection

SAT **:** Speaker Adaptive Training

SI-SDR **:** Scale-Invariant Signal-to-Distortion Ratio

SOTA **:** State-of-the-art

STOI **:** Short-Time Objective Intelligibility

STT **:** Speech-to-Text

TTS **:** Text-to-Speech

VAD **:** Voice Activity Detection

WAV **:** Waveform Audio

WER **:** Word Error Rate

WSJ **:** Wall Street Journal

WebRTC **:** Web Real-Time Communications

G2P **:** Grapheme-to-Phoneme

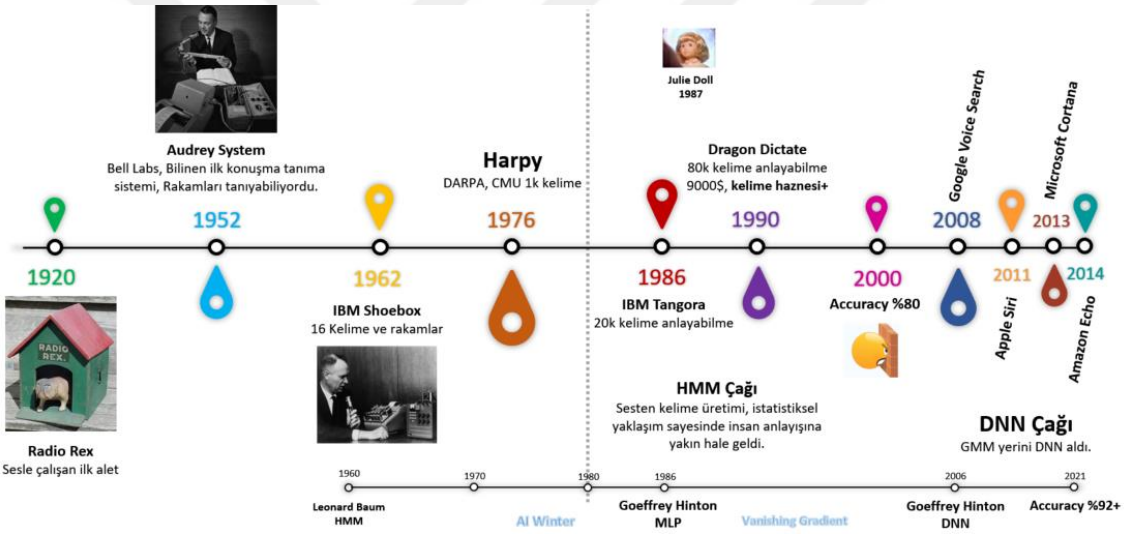
ŞEKİL LİSTESİ

TABLO LİSTESİ

1. GİRİŞ

1.1. Konuşma Tanımanın Geçmişi ve Bugünü

Konuşma tanıma günümüzde popüler olarak kullanılan bir teknoloji olmakta ve insan-makine etkileşimi için olmazsa olmaz köşe taşı konumunda bulunmaktadır. Her yıl gelişmekte olan konuşma tanıma teknolojisi 80'li yıllarda ilk sıçramasını, Artificial Neural Networks (ANN) ile ikinci sıçramasını gerçekleştirmiştir. ANN öncesi ise geleneksel yöntemlerin çağı olarak düşünülebilir. Bu dönemlerde konuşma tanımanın istatistiksel çözümleri üzerinde durulmuştur. Hidden Markov Model (HMM) Bayes teoremini temel alarak akustik modellemede öne çıkmıştır. Basit olarak bu sistemler giriş sinyali için fonem olasılıklarını alarak, bu olasılıkları bir Gaussian Mixture Model (GMM) ile sınıflandırarak çalışmaktadır.



Şekil 1.1: Konuşmanın geçmişi ve kırılma noktaları.

Şekil 1.1'de görüldüğü gibi ilk kırılım 80'li yıllarda istatistiksel yöntemlerin Large Vocabulary Continious Speech Recognition (LVCSR) sistemleri mümkün kılması ile oluşmuştur. Bu kısaltma 5000 ile 60.000 arasında kelimeyi başarılı şekilde sınıflandırabilen sistemler için kullanılmıştır [1]. Sınıflandırabildikleri kelime sayısının yanında sürekli konuşma tanıma yapabilmeleri de bir yenilik olmuştur. Zira 80'ler öncesinde dek konuşma tanıma, duraksamalı ve daha çok komut yöntemi ile çalışan bir teknoloji olmuştur. Örneğin, 1952'de ilk çalışmaları yapılan konuşma tanıma sistemleri yalnızca 10 kelimeyi başarıyla sınıflandırabilmekteydi. Sonrasında Sovyet araştırmacıların Dynamic Time Warping (DTW) çalışmasını konuşma tanımada ilk

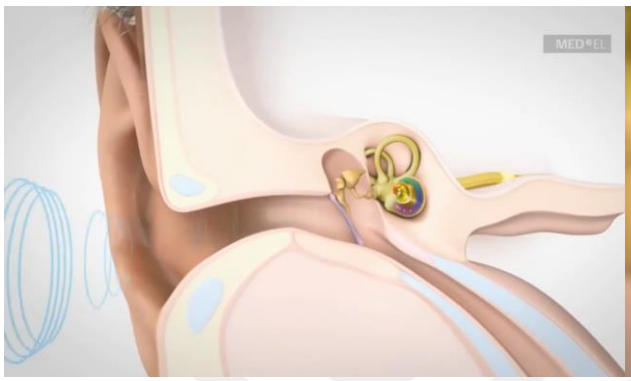
olarak Raj Reddy kullanarak başarılı bir araştırmayı Carnegie Mellon Üniversitesi'nde gerçekleştirdi. Bu sistem 200 kelime tanıma kabiliyetine sahipti. 70'lerde IBM 'Tangora' isimli sistem üzerinde çalışmaya başladı ve bu sistem 80'ler ortalarında 20.000 kelime tanıyabilmekteydi. Fakat bu sistemler duraksamalı telaffuza gereksinim duymaktaydı. 80'ler konuşma tanıma problemin HMM modellerinin altın çağı olarak yorumlandı. Bu zamanlarda probleme ilk kez istatistiksel yaklaşıldı. İlk başarılı sürekli konuşma tanıma örnekleri Defence Advanced Research Projects Community (DARPA) tarafından desteklenen araştırmalarla geldi ve bu sistemler 1000 kelimeyi kesintisiz bir konuşmada yüksek başarıyla tespit edebilmekteydi [2].

İstatistiksel yaklaşımı esas alan HMM 80'li yıllar ve sonrasında başarılı akustik modellemenin vazgeçilmez teknolojisi olmuştur. Öte yandan Sinir Ağları üzerinde de çalışmalar gerçekleştirilmiştir. Bu yıllarda Sinir Ağları için "Vanishing Gradient" (Kaybolan Gradyan) probleminin çözümü bu teknolojinin kullanılmasının önündeki en büyük engel olmuştur [2]. Bu problemin aktivasyon fonksiyonlarının ele alınmasıyla çözülmesi sonrasında Microsoft Research katkısıyla Hinton ve arkadaşları [3]'de ilk kez Sinir Ağları yardımıyla akustik modelleme üzerinde çalışmışlardır. Ardından 97'de Hochreiter ve Schmidhuber'in tanıttığı Long-Short Term Memory (LSTM) [4] eğitimi 2006'da Graves ve arkadaşları tarafından tanıtılan Connectionist Temporal Classification (CTC) ile mümkün olmuştur. [5]. Fakat RNN eğitimi normal normal Sinir Ağlarından farklı olarak daha zorlu olduğundan pratik kullanımına dair örnekler nispeten geç gelmiştir. Google 2015'te LSTM-CTC tabanlı bir konuşma tanıma altyapısı sunarak teknolojiyi büyük ölçüde etkilemiştir [6]. Bu dönemlerde ASR farklı diller için uygulanırken sürekli olarak dile bağlı fonetik bilgi birikimi gereksinimine ihtiyaç duymaktaydı. Telaffuz için akustik modeli besleyecek Lexicon sözlüğü, Akustik model için en iyi öznitelikle çalışabilecek model seçimi bunun yanında dile özgü kelimelerin birleşimini ve kurallarını öngören veya düzelten dil modelleri olasılıksal olarak birbirlerine bağlanmaktaydı. Bunun yanında tek modeli esas alan uçtan-uca ise 2014'te gerçekleştirildi [7].

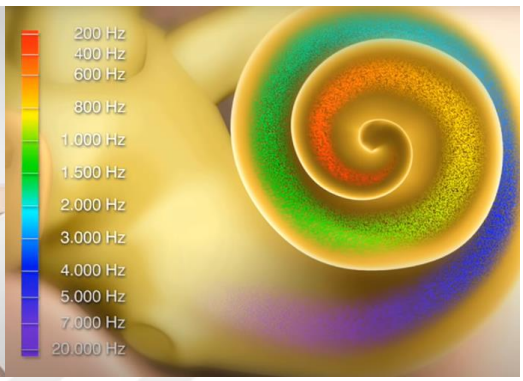
Konuşma tanımanın bugünü ise ikinci kırılım sonrası başlamaktadır. Çünkü uçtan-uca modeller eski yapıları hem kökten değiştirmeyi önermiş hem de umut verici sonuçlar ortaya koymaya başlamıştır. Uçtan-uca modeller için bir sonraki adım ise büyük ölçekli veri setleri ile yapılan çalışmalarda olacaktır.

1.2. İnsan Duyma Algısı

İnsan kulağının algılayabildiği ve algılayamadığı ses frekansları bulunmaktadır. Bir konuşma tanıma sisteminin insan algısını taklit edebilmesi için bu frekanslara dikkat etmesi önemli olmaktadır. Şekil 1.2'de insan kulağının iç yapısı ve sesi algılayan cochlea yapısı ise Şekil 1.3'te verilmiştir.



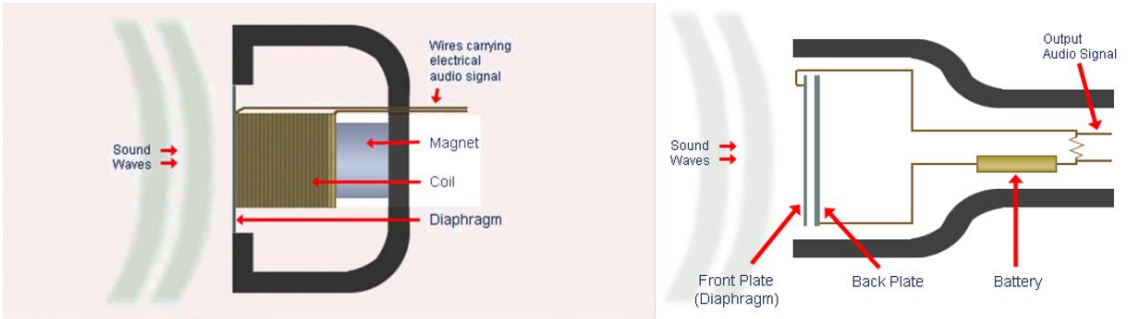
Şekil 1.2: İnsan kulağının iç yapısı [8].



Şekil 1.3: İnsan kulak cochleasına ait frekans reseptör dağılımı [8].

1.3. Dijital Sesin Oluşumu ve Saklanması

Ses dalgaları mikrofonlar aracılığıyla sayısallaştırılmaktadır. Mikrofonlar kendine ulaşan ses dalgalarını hassas yapıları sayesinde bir modülasyona sokarak sayısal sesi oluştururlar. Temelde oluşturulan sayısal veriler, mikrofonun iç mekanizmasının hareketi veya titreşimiyle oluşmaktadır. Mikrofonların iç yapısı Şekil 1.4'te gösterilmiştir.



Şekil 1.4: Mikrofonların çalışma prensibi.

Sayısallaştırılan ses istenir bir takım sıkıştırma algoritmaları ile küçültülerek istenirse de sıkıştırma yapılmadan, ham halde saklanmaktadır. Web için çoğunlukla .mp3 ve .m4a sıkıştırılmış ses formatları tercih edilirken çoğu konuşma veri setinde sıkıştırmanın veriyi bozmaması adına .wav uzantısı tercih edilmektedir.

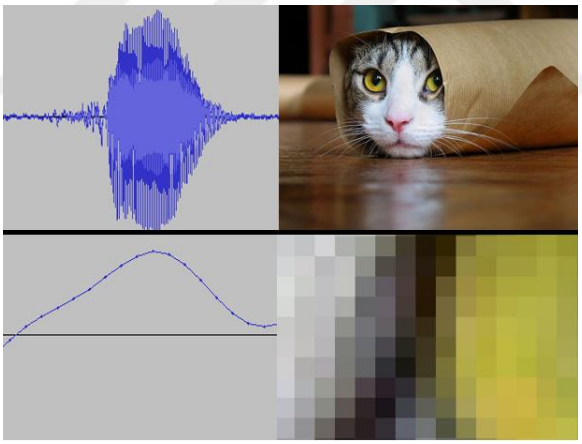
1.4. Dijital Ses ve Görüntü Arasındaki Benzerlikler

Dijital ses ile görüntü arasında benzerlikler bulunmaktadır. Bu benzerlikler ses verine ait parametrelerin iyi anlaşılmasını kolaylaştırabilir. Bu benzerlikler Tablo 1.1'de verilmiştir.

Tablo 1.1: Dijital ses ve görüntü arasındaki benzerlikler.

Ses	Görüntü
Frekans	Orijinal Çözünürlük
Örnek başına bit değeri	
Ses kanalı (Mono, Stereo..)	Görüntü Kanalı (R, G, B, A..)
Süre	En ve boy

Görüntüler için her zaman pikseller, ses verileri için ise her zaman örnekler en küçük bilgiyi taşırlar. Her iki durumda da veri bit düzeyinde temsil edilmektedir. Şekil 1.5'te ses ve görüntüye dair daha detaylı bir görsel verilmiştir. Yakında anlamsız gelen bilgi daha uzaktan anlamlı bir görsele dönüşürken, sesin dinlenilmesi gerekmektedir.



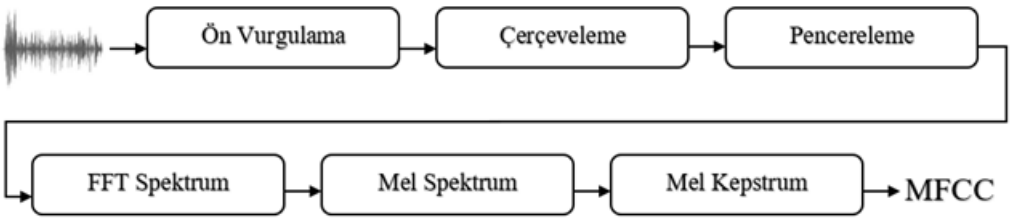
Şekil 1.5: Dijital ses ve görüntünün daha yakın incelenmesi.

1.5. Konuşma Öznitelikleri

Öznitelik kalitesi konuşma tanımada kullanılan model başarıları için kritik öneme sahip olmaktadır. Ses zamanla değişen bir yapıya sahip olduğundan çoğu ses tanıma yöntemi sinyalleri parçalara bölerek işlemeyi taban alır. Öznitelikler için farklı yöntemler bulunmaktadır; Mel Frequency Cepstrum (MFC), Linear Predictive Codes (LPC), Perceptual Linear Prediction (PLP) bunlardan birkaçıdır [9]. Bu çalışmadaki modellerde MFC öznitelikleri kullanılmıştır.

1.5.1. MFCC

Mel ölçeği insan duyma algısına göre tasarlanmıştır. Ses algımız, 1kHz'e kadar sürekli ve doğrusal sonrasında ise logaritmik artışla bir ses algısına sahiptir. Mel Frequency Cepstral Coefficient (MFCC) öznitelik çıkarımı Şekil 1.6'daki gibi özetlenebilmektedir.



Şekil 1.6: MFCC adımlarına ait blok diyagramı.

1.5.1.1. Ön Vurgulama

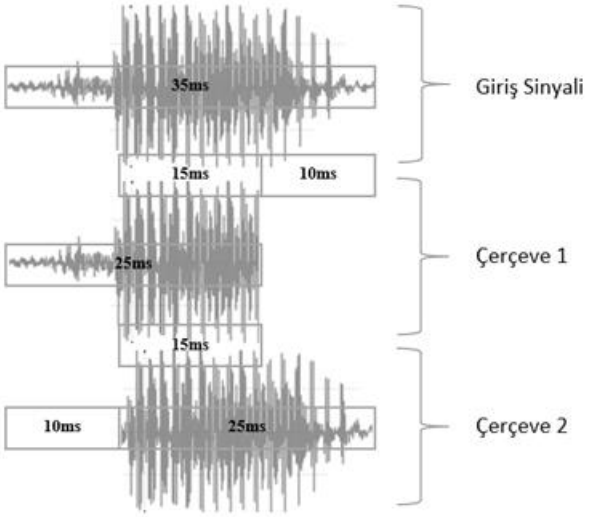
Bu aşama frekans uzayının yumuşatılması için kullanılmaktadır. Amacı, "insan ses üretim mekanizması sırasında bastırılan konuşma sinyalinin yüksek frekanslı kısmını telafi etmektir" [10]. Ön vurgulama bağıntı (1.1) ile gerçekleştirilmektedir.

$$Y[n] = X[n] - \alpha * X[n-1] \tag{1.1}$$

Burada X, Y, α sırasıyla; sinyal, vurgulanmış sinyal, vurgulama katsayısını ifade etmektedir. Genelde α için 0,95 veya 0,97 tercih edilir. Girdi ve çıktı sinyal uzunlukları aynı ve formül gereği n = 1, 2, 3… olarak alınmaktadır.

1.5.1.2. Çerçeveleme

Sinyalin parçalanması için kullanılmaktadır. "Framing" olarak da isimlendirilmektedir. Bu aşamada, ses analizinin vazgeçilmez bir parçası olarak belirli genişlikte bir çerçeve sinyal üzerinden kaydırılarak, sinyalden alt parçalar alınmaktadır. Kayma miktarı çerçeve genişliğini geçmeyecek şekilde seçilmesi gerektiğinden çerçeveler bir önceki çerçeve ile kısmen örtüşmektedir. Örtüşmenin miktarı analizin daha keskinliğini artırırken, hesaplamayı yavaşlatmaktadır. Burada çerçeve genişliği 20-25ms, adım uzunluğu ise 10-15ms alınmaktadır. Çerçeveleme Şekil 1.7'deki görsel ile anlatılmıştır.

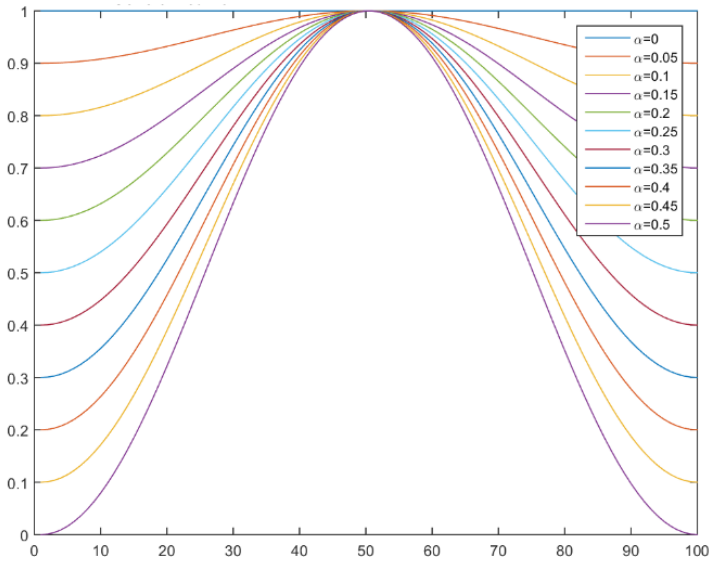


Şekil 1.7: Çerçevelemenin 35ms sinyal üzerindeki gösterimi.

1.5.1.3. Pencereleme

Ses analizi için genel yaklaşım frekans uzayına geçip öznitelikleri buradan çıkarmaktır. Frekans uzayına geçmeden önce FFT uygulamasına yardımcı olması açısından sinyale bir başka fonksiyon ile pencereleme uygulanmaktadır. Bu aşamanın temel amacı FFT uygulanacak sinyali sürekli bir sinyal haline getirebilmektir. Pencereleme aşaması için genellikle (1.2) bağıntısındaki Hamming veya Hanning penceresi kullanılmaktadır.

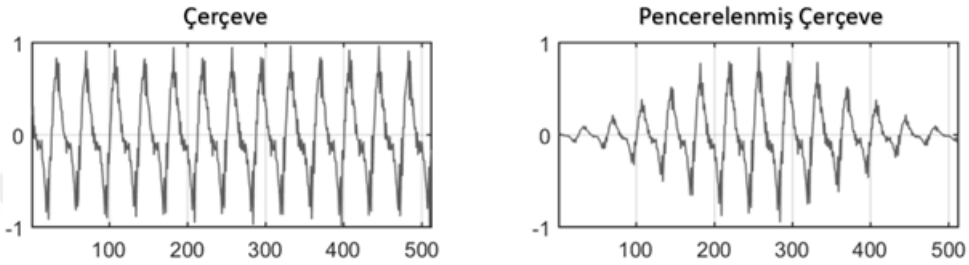
$$w[n] = (1-\alpha) - \alpha * \cos\left(\frac{2\pi n}{N-1}\right), \quad 0 \leq n \leq N-1 \tag{1.2}$$



Şekil 1.8: Farklı katsayılarının oluşturduğu Hamming pencereleri.

Burada N giriş için örnek miktarı, α genelde 0,46 olarak alınan bir yumuşatma katsayısı olduğundan denklem $w[n] = 0.54 - 0.46 * \cos\left(\frac{2\pi n}{N-1}\right)$ halini almaktadır. Farklı değerler için pencereleme fonksiyonunun temsilleri Şekil 1.8'de gösterilmiştir. Bu fonksiyon sinyal boyunca bağıntı (1.3) gibi uygulanarak pencereleme aşaması tamamlanmaktadır.

$$Y[n] = X[n] * w[n] \tag{1.3}$$



Şekil 1.9: Denklem (1.3) ile pencerelenmiş bir çerçeve.

Şekil 1.9'da pencereleme yapılmıştır bir sinyal gösterilmektedir. Etkinin görselleşmesi için çerçeveler uzun tutulmuştur. Normalde çerçeve genişlikleri daha kısa olmaktadır. Dikkat dikkat uyandıran nokta, pencere uygulanmış sinyaldeki başlangıç ve bitiş noktalarının aynı seviyede kalması olmaktadır. Sürekli sinyallerin bir özelliği ise bu şekilde sağlanarak FFT spektrum uzayı yumuşatılmaktadır.

1.5.1.4. FFT Spektrum

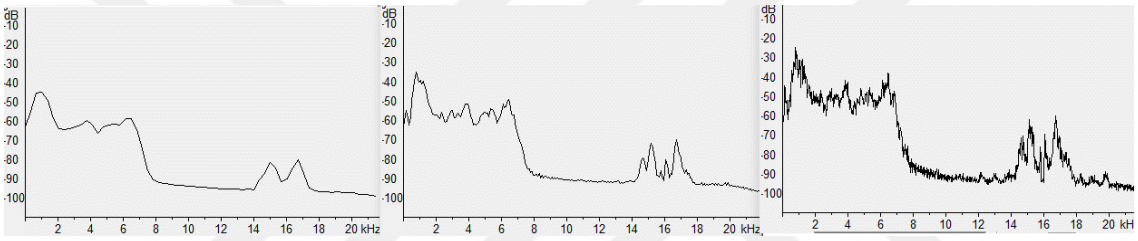
Spektral analiz veya spektrum analizi, bilgilerin incelenmesi ve sınıflandırılması için sıkça kullanılan yöntemler arasında yer almaktadır. Elektromanyetik spektrum, frekans spektrumu, enerji spektrumu veya Mel Spektrumu gibi farklı spektrum türleri bulunmaktadır. Bu aşama ses analizi özellikle frekans alanına geçmek amacıyla sıkça tercih edilmektedir. Bu tercihin arkasındaki nedenlerden biri, sesin temel özelliklerinin bu analiz yöntemiyle daha açık bir şekilde tespit edilebilmesi olmaktadır. Ayrıca, kaynaktan alınan sesin bazen parazit içerebilmesi ve ana bilginin berraklığını bozabilmesi nedeniyle frekans analizi tercih edilmektedir.

Birçok öznitelik araştırması genellikle Frekans spektrumunun oluşturulmasıyla başlamaktadır. Bu süreçte, sinyal genellikle sinüs dalgaları kullanılarak oluşturulur. Örneğin, konuşma sinyalleri ses tellerinin titreşimlerinden kaynaklanır ve zaman bazen bu tür sinyaller için önemli değildir. Asıl önemli olan, bu sinyalin oluşturduğu frekanstır

[11]. Dijital ses sinyalleri kaynaktan örnekleme yapıldığından, bu tür bilgileri çıkarmak için ifade (1.4) Discrete Fourier Transform (DFT) kullanılmaktadır.

$$X_k = \sum_{n=0}^{N-1} x_n e^{-i2\pi kn/N}, k = 0, \dots, N-1 \tag{1.4}$$

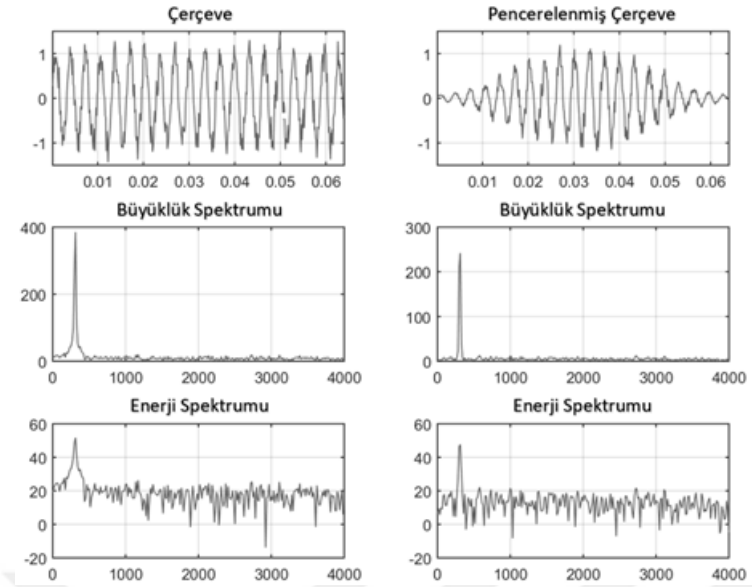
Aynı sinyale uygulanacak DFT, $O(N^2)$ karmaşıklığa sahipken FFT algoritması yakın veya aynı analizi $O(N \log N)$ işlem karmaşıklığında ile gerçekleştirilmektedir. Bu nedenle uygulama noktasında FFT tercih edilmektedir. FFT genelde 512 veya 256 nokta üzerinden hesaplanmaktadır. Bu miktarın yükselmesi hesap süresini artırabilirken az oluşu da analizin kesinliğini etkilemektedir. Şekil 1.10'da bu farklar gösterilmektedir.



Şekil 1.10: Soldan sağa sırasıyla 128, 512, 2048 nokta FFT analizlerinin enerji spektrumu.

FFT her sonuç çerçevesi için uygulanmasının ardından kutupsal koordinat sonuçlarından büyüklük bileşeni $|FFT(x_i)| = \sqrt{R^2 + I^2}$ çekilmektedir. Büyüklük bilgilerinden güç hesaplanarak bu noktada güç spektrumu elde edilmektedir. Bu işlem N adet FFT noktası kullanılarak $P = \frac{|FFT(x_i)|^2}{N}$ ifadesiyle yapılmaktadır. Burada önemli bir nokta ise Hamming pencerelerinin yaptığı etki olmaktadır.

Fourier dönüşümü, periyodik bir sinyale uygulandığında sonuç, frekans spektrumunu temsil eden ayrık bir grafik oluşturur. Periyodik olmayan sinyaller için ise frekans spektrumu sürekli bir grafikle temsil edilir. Bu farklılık, sinüzoidal sinyallerin periyodik sinyallerde daha kolayca temsil edilebilmesinden kaynaklanır. Bu nedenle, sinyali sürekli hale getirmek amaçlanır, bu da sonuçta elde edilecek FFT sonuçlarının daha keskin ve daha kolay okunabilir hale gelmesini sağlar. Şekil 1.11, bu etkiyi göstermektedir.



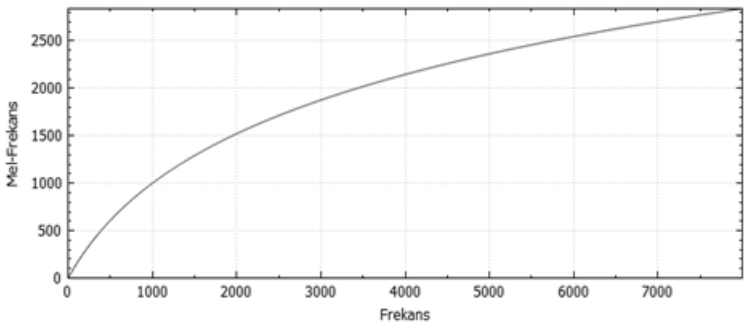
Şekil 1.11: Hamming penceresi uygulanmış (sağ) uygulanmamış (sol) sonuçlar.

1.5.1.5. Mel Spektrum

Mel birimi, insan kulağını taklit etmek amacıyla tasarlanmış bir birimdir. Yani Mel birimi, ses frekanslarını insan kulağının nasıl algıladığına dayalı olarak oluşturulmuştur ve frekansı doğrusal bir eksen yerine insan algısına daha uygun bir şekilde temsil etmektedir [12]. Bağıntı (1.5), frekansı Mel-frekansa dönüştürmek için kullanılırken, bağıntı (1.6) ise Mel-frekansını frekansa döndürmek için kullanılmaktadır. Bağıntı (1.5) kullanılarak elde edilen grafik, normal frekans ve Mel-frekans arasındaki ilişkiyi gözlemlememize yardımcı olmaktadır.

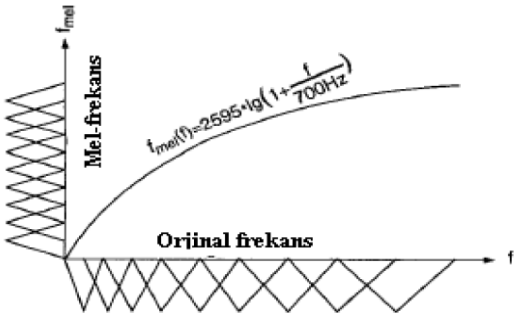
$$m(f) = 2595 * \log 10\left(1 + \frac{f}{700}\right) \tag{1.5}$$

$$f(m) = 700\left(10^{\frac{m}{2595}} - 1\right) \tag{1.6}$$



Şekil 1.12: Frekans-Mel Frekans grafiği.

Şekil 1.12'de frekans ve Mel frekansı arasındaki farkı gözlemlenebilir. Dikkat edilirse, 1 kHz'e kadar olan bölgede Mel ve normal frekanslar neredeyse lineer bir artış gösterirken, bu noktadan sonra Mel-frekansı logaritmik bir artış sergilemektedir. Yapılan araştırmalar, insan kulağının algısının da bu şekilde olduğunu göstermektedir. MFC (Mel-Frequency Cepstral Coefficients) elde edilirken, bu aşamada üçgen filtre setleri oluşturulur. Bu filtre setleri, Mel ölçeği için sabit aralıklarla konumlandırıldığından, normal frekanslar açısından bakıldığında bir tür ötelenmiş izlenimi verebilir. Şekil 1.13, bu durumu net bir şekilde göstermektedir.



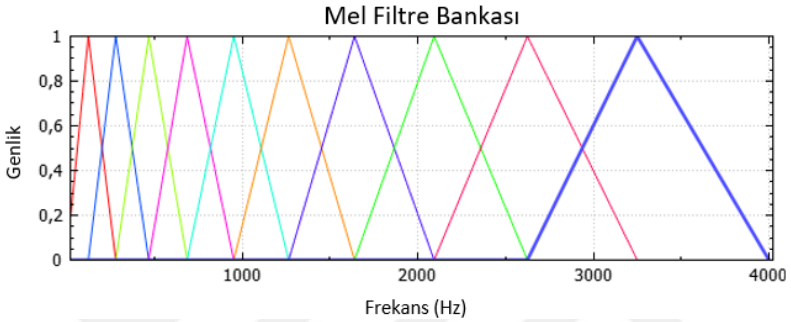
Şekil 1.13: Mel-filtreleri ve frekans üzerinde temsilleri.

Filtre sayısı tercihlere bağlı olarak genellikle 26 ile 40 arasında değişebilir, ancak 40 filtre sıklıkla kullanılmaktadır. Bu filtreler, ses verisinin örnekleme oranının yarısı temel alınarak oluşturulur. Bunun nedeni, verinin Nyquist örnekleme kuralına göre alındığı varsayılırsa, içerdiği bilginin maksimum frekansının örnekleme oranının en fazla yarısına eşit olması gerektiğidir. Bu filtreler, bağıntı kümesi (1.7) kullanılarak oluşturulmaktadır.

$$H_m(k) = \begin{cases} 0 & k < f(m-1) \\ \dfrac{k - f(m-1)}{f(m) - f(m-1)} & f(m-1) \le k < f(m) \\ 1 & k = f(m) \\ \dfrac{f(m+1) - k}{f(m+1) - f(m)} & f(m) < k \le f(m+1) \\ 0 & k > f(m-1) \end{cases} \tag{1.7}$$

Burada m filtre sayısını, k filtre içindeki noktaları, f her bir filtreye karşılık gelen FFT orta noktasını (bunlara FFTbins adı da verilir), H_m üçgen filtreyi temsil eder. Örnek olarak 8000Hz örnekleme oranına sahip bir sinyal için filtre bankası şu şekilde kurulur;

- $m = 10, FFT_s = 512$ *olsun.*
- $f_{ö} = 8000$ *ise* $f_{alt} = 0, f_{üst} = 8000 / 2.$
- *Bağıntı (1.5) kullanılarak* $m(f_{alt}) = 0.0$, $m(f_{üst}) = 2146.06$ *hesaplanır. Bu Mel aralığı daha önce de belirtildiği gibi lineer olarak artış sağlanması için* $m - 1$ *parçaya bölünür. Alt ve üst değerler dahil olmak üzere buradan 12 nokta elde edilir.*
- *Bu 12 noktadan bağıntı (1.6) ile frekans uzayına geçilir.*
- *Her bir noktanın FFT sonuçlarındaki yeri bulunur.* $f(i) = (FFT_s + 1) * \frac{h(i)}{f_ö}$ *burada* $h(i)$ *12 Noktayı* $f(i)$ *ise bağıntı (1.7)'deki* f *'yi temsil etmektedir. Şekil 1.14 burada bahsedilen filtre bankasını göstermektedir.*

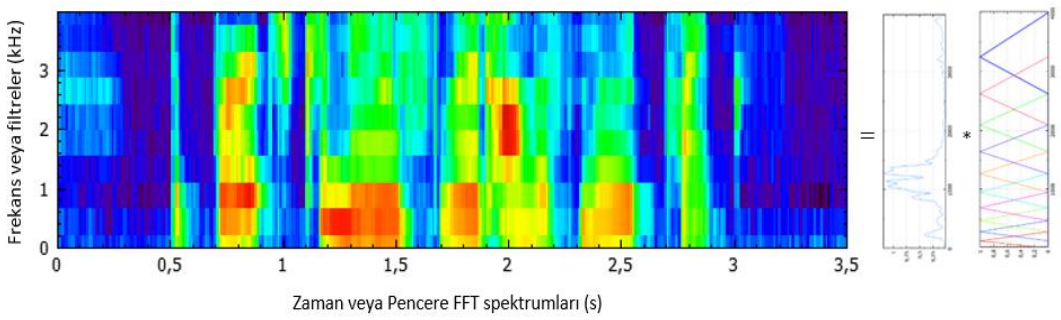


Şekil 1.14: 10 Adet Mel filtresi. 10. Filtre, yapısı gösterilmesi amacıyla seçili haldedir.

Oluşturulan 10 üçgen filtre Şekil 1.14'te gösterilmiştir. Mel spektrumuna geçişte üçgen geçişli filtrelerin kullanılmasının nedeni, büyüklük spektrumunun pürüzsüzleştirilmesi ve elde edilecek özniteliklerin azaltılmasıdır [10]. Bu konu, bir sonraki bölümde spektrogramlar üzerinde daha açık bir şekilde gözlemlenecektir.

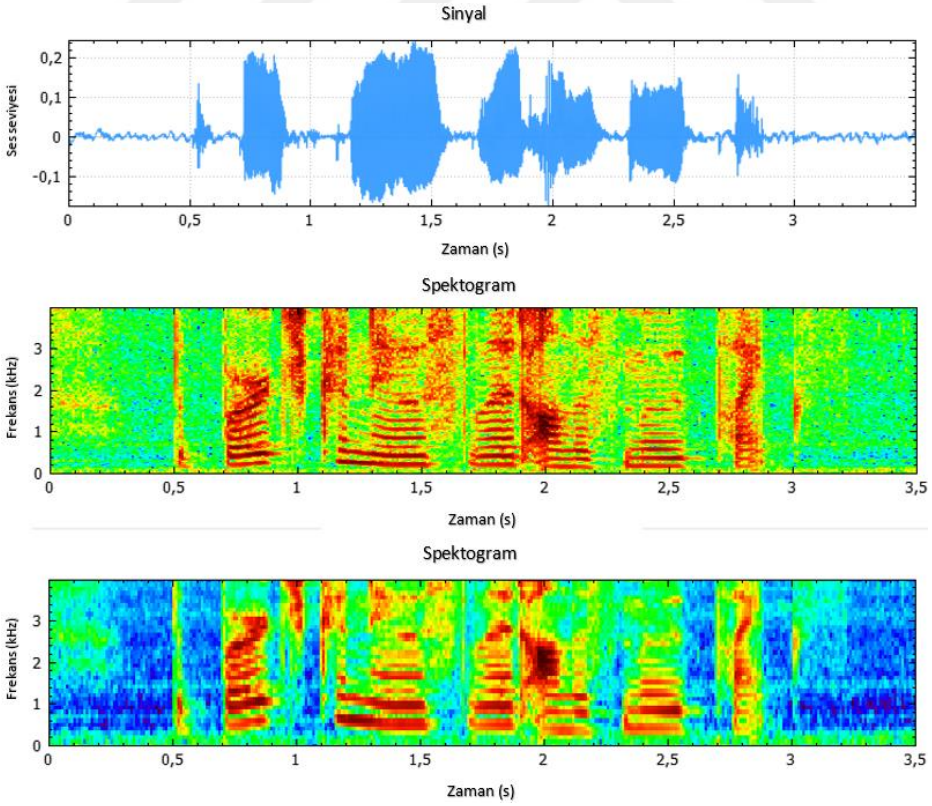
1.5.1.6. Mel Kepstrum

Tüm analiz pencereleri FTT yöntemi kullanılarak frekans bileşenlerine dönüştürülerek, spektrogram olarak adlandırılan bir gösterim elde edilir. Spektrogram, zamanı yatay eksende, frekansı dikey eksende ve bu iki eksenin kesişim noktasında belirli bir frekanstaki genliği renklerle ifade eden üç boyutlu bir gösterim şeklidir [9]. Mel spektrumunu oluşturmak için, her nokta filtredeki Güç spektrumundaki karşılığı ile çarpılır ve sonuçlar toplanır. Bu hesaplama sonucunda, ilgili filtrenin belirli bir çerçeve anındaki Mel enerjisi elde edilir ve bu enerjinin doğal logaritması alınır. Bu işlem tüm çerçeve anları için tekrarlandığında, "spektrogram" adını verdiğimiz geniş analiz grafiği oluşturulur. Spektrogram, spektrumların birleşiminden oluşan üç boyutlu bir grafiktir. Şekil 1.14, filtre bankası kullanılarak oluşturulan spektrogramı gösterir ve bu işlem Şekil 1.15'te açıklanmaya çalışılmıştır.



Şekil 1.15: 3,5 saniye süren sinyal üzerinde 10 Mel-filtresi kullanılarak oluşturulmuş spektrogram.

Mel filtrelerinin enerji toplamlarının logaritmasının alınmasının nedeni, frekans analizlerinden elde edilecek sonuç bilgilerinin çok küçük değerlere sahip olabileceğidir. Bu tür durumlarda, bilgiyi daha görünür hale getirmenin bir yolu olarak logaritma fonksiyonları kullanılmaktadır. Bu işlem için doğal logaritma veya 20-30 katsayıları ile çarpılarak onluk tabanda logaritma alma gibi teknikler sıklıkla kullanılmaktadır. Rastgele bir konuşma sinyali üzerinde 40 filtre kullanılarak oluşturulan Mel ve normalize edilmiş FFT spektrogramları Şekil 1.16'da gösterilmiştir.



Şekil 1.16: En üstte girdi sinyali. Ortada bu sinyale ait FFT spektrogram. En altta 40 Mel filtresi kullanılarak oluşturulmuş spektrogram.

Kepstrum (Cepstrum), bir sinyalin tahmini spektrumunun logaritmasının Inverse Fourier Transform (IFT) işlemine tabi tutulması sonucu elde edilen bir analiz yöntemidir. Güç spektrumundan türetilen bir türüne Güç Kepstrum denilmektedir. Ses analizinde, bu aşama Discrete Cosine Transform (DCT) aracılığıyla gerçekleştirilmektedir. Bu yöntem, frekans uzayından zaman uzayına geçiş yapmayı sağlar. Kepstrum analizi, özellikle ses işleme ve konuşma tanıma gibi alanlarda kullanılır ve ses sinyalinin özelliklerini çıkarmak için faydalı bir araç olmaktadır.

$$C_m = \sum_{k=1}^{N} \cos[m * (k - 0.5) * \pi / N] * E_k, \qquad m = 1,2, \dots, L \tag{1.8}$$

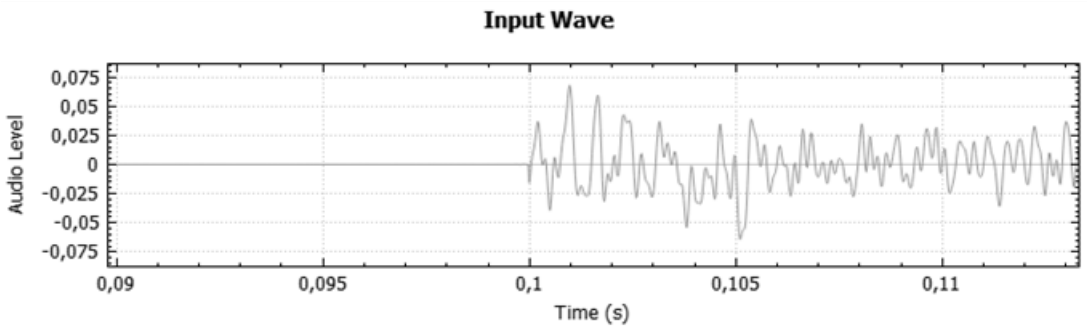
Burada elde edilmek istenen katsayı sayısını, N üçgen filtre setini, spektrogramdaki logaritmik enerji sayılarını göstermektedir. Genellikle filtre sayısı 26 ila 40 arasında değişirken, katsayı sayısı genellikle 12 olarak seçilmektedir. İlk katsayı genellikle ortalama logaritmik enerjiyi temsil ettiği için hesaplamalarda dikkate alınmamaktadır. Elde edilen her katsayı, analiz penceresinin boyutuna göre normalize edilir ve bu katsayılar MFCC katsayıları olarak adlandırılırlar. DCT için kullanılan (1.8) ifadesindeki sonuçlar bazı yazarlar tarafından DCT matrisinin ortogonal hale getirilmesi için $\sqrt{\frac{2}{N}}$ ile çarpılır. Sinyali daha iyi temsil için enerji, sıfır geçiş, delta, delta-delta gibi farklı öznitelikler de eklenebilir.

1.5.2. Enerji ve Sıfır Geçiş

Enerji özelliği, sinyalin hiçbir işleme tabi tutulmadan, ham haliyle çerçeveler aracılığıyla elde edilir. Sinyalin gücü, frekans uzayındaki güçle karıştırılmamalıdır. Ayrı bir sinyalin enerjisi, (1.9) denklemi kullanılarak hesaplanır. Alternatif olarak, sinyalin ortalama gücü de kullanılabilir, bu durumda enerji uzunluğa bölünür.

$$E = \sum_{n=-\infty}^{\infty} |X[n]|^2 \tag{1.9}$$

Burada E çerçeve enerjisini, X[n] sinyal örneğini temsil eder. Sıfır geçiş özelliği de konuşmalar için belirleyici bir özelliktir. Sessizlik anında sıfır geçişi keskin değil ve az iken konuşma anlarında sıfır geçişleri hem keskin hem de fazladır. Bu durum şekil 1.17'de rahatlıkla gözlenebilir.



Şekil 1.17: Sıfır geçiş özniteliğinin 0.1. saniyede konuşmanın başlamasıyla artışı gözlenebilir.

Sıfır geçişi hesaplamak için ifade (1.10) kullanılabilir. ZCR Sıfır geçiş sayısını, sgn işaret fonksiyonunu, N sinyal uzunluğunu ve X sinyali temsil etmektedir.

$$ZCR = \frac{1}{2N}\sum_{n=1}^{\infty}|sgn(X[n]) - sgn(X[n-1])| \tag{1.10}$$

Bu öznitelik, sinyal işlem görmeden önce hesaplanmalı ve büyük sayılar üretebileceği için normalize edilmesi gerekmektedir. Sıfır geçiş özelliği hesaplanırken türevler de kullanılabilir. Bu şekilde, türev büyüklüklerine bakılarak sessizlik anındaki değişiklikler daha az veya hiç dikkate alınabilir.

1.6. Akustik Modelleme

İyi öznitelikler beraberinde iyi temsilleri getirirken, bu temsillerin sınıflandırma öncesi probleme özgü modellenmesi başarıyı artırabilmektedir. Özellikle geleneksel yöntemlerin vazgeçilmesi olan akustik modellemede dile özgü fonem sunumları kullanılmaktadır. Bazı fonetik olmayan diller için özellikle dildeki harf miktarından daha fazla fonem bulunurken Türkçe gibi fonetik dillerde fonem sunumlarını oluşturmak alfabedeki harfleri kullanmak kadar kolay olabilmektedir. Tablo 1.2'de İngilizce ve Türkçe örnek fonem setleri gösterilmiştir (fonemler birbirlerinin eşdeğeri değildir). Türkçe fonem seti Tablo 1.2'de gösterilmesinin yanında daha farklı fonetik sesleri kapsayacak şekilde yeniden oluşturulabilir. Fonemlerin konuşma tanımaya katkısı ilgili konuşma dilindeki farklı tınıları temsil edebilme başarılarıdır olmaktadır. Sonrasında bu setler akustik model oluşumunda ve nihai çeviri oluşumunda kullanılmaktadır [1].

Tablo 1.2: Örnek İngilizce ve Türkçe fonem setleri.

#	İngilizce	Türkçe	#	İngilizce	Türkçe
1	AA	A	**21**	L	R
2	AE	B	**22**	M	S
3	AH	C	**23**	N	Ş
4	AO	Ç	**24**	NG	T
5	AW	D	**25**	OW	U
6	AY	E	**26**	OY	Ü
7	B	F	**27**	P	V
8	CH	G	**28**	R	Y
9	D	Ğ	**29**	S	Z
10	DH	H	**30**	SH	-
11	EH	I	**31**	T	-
12	ER	İ	**32**	TH	-
13	EY	J	**33**	UH	-
14	F	K	**34**	UW	-
15	G	L	**35**	V	-
16	HH	M	**36**	W	-
17	IH	N	**37**	Y	-
18	IY	O	**38**	Z	-
19	JH	Ö	**39**	ZH	-
20	K	P			

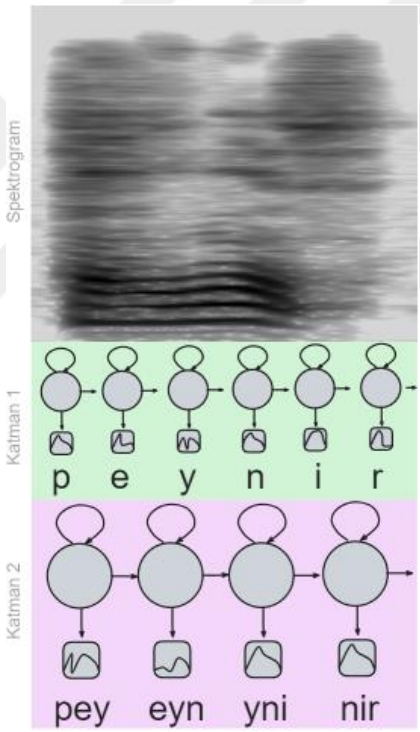
Akustik modellemeye sıklıkla telaffuz sözlüğü adı verilen ve lexicon olarak belirtilen sözlük ile başlanmaktadır. Veri içindeki her kelime karşılığı olan fonem açılımlarına ayrıştırılmalıdır. Literatürde bu aşamaya grapheme-to-phoneme (g2p) adı verilmektedir [1]. Tablo 1.3'te bir örnek g2p sürecini temsilen verilmiştir (Türkçe fonem seti alana veya dile özgü tınıları kapsayacak şekilde tekrar düzenlenebilir).

Türkçe g2p aşaması için büyük bir zorluk barındırmaktadır. Bunun sebebi ise Türkçenin eklemeli bir dil oluşu ve çok fazla farklı kelime üretebilme beceresine sahip olmasıdır. Örnek olarak ortalama bir Türkçe lexicon sözlük rahatlıkla 1,5 milyon kelimeye erişebilirken, İngilizce bir lexicon için 400 bin kelime yeterli gelebilmektedir.

Tablo 1.3: Örnek g2p gösterimi.

Kelime	Açılım	Kelime	Açılım
cheese	CH IY Z	**peynir**	P E Y N İ R
hurt	HH ER T	**acı**	A C I
fee	F IY	**ücret**	Ü C R E T
green	G R IY N	**yeşil**	Y E Ş İ L

Geleneksel yaklaşımda sıklıkla akustik modelleme için HMM kullanılmaktadır. Bu yapı yenilikçi yöntemlerde yerini tamamen yenilikçi sinir ağlarına bırakmıştır [5-8]. Örnek bir fonem modeli Şekil 1.18'de gösterilmiştir.



Şekil 1.18: Peynir söylemi için HMM modeli.

Peynir söylemi için gösterimi yapılan HMM modelinde katman 1'de fonem çıkarımı yapılırken, katman 2'de ise bu fonemlerin birleşimlerinden farklı bir modelleme gerçeklenmiştir. Akustik modelleme bu yöntemin kaç farklı şekilde tekrar edeceği ve modelleme içerisine ne gibi dilsel özellikler ekleneceğiyle ilgilenmektedir.

1.7. Dil Modelleri

Dil modelleri, konuşma tanımada sentezin başarısını artırmada kullanılmaktadır. Her dilin kendine ait bir yapısı olduğundan, örnek olarak; Türkçe'de özne + nesne + yüklem düzeni veya İngilizce'de özne + yüklem + nesne düzeninden yola çıkılarak, sentezlenen cümlede bulunan sorunlar tespit edilip, ekleme ve çıkarma yapılabilmektedir. Sıklıkla dil modelleri çok büyük yazı derlemleri (corpus) aracılığıyla oluşturulmaktadır. Amaç yöntemin olasılıksal olarak doygunluğa ulaşabilmesi ve kesinliğin artırılmasıdır [1, 10].

En yaygın metot olarak n-gram yaklaşımını tercih edilmektedir. Burada 'n' peşi sıra kaç kelimenin birbiriyle ilişkilendirileceğini belirlemektedir. Değer arttıkça hesaplama güçleşmektedir. Her kelimenin n belirli kelime ile sıralanmasında olasılıkların en yükseği muhtemel sonuç olmaktadır. Tablo 1.4'te "peynirli ve sucuklu tost sipariş ettim" çıktısının beklendiğinde konuşma tanıma sonucu sentezlenen cümledeki farklı olası hatalar gösterilmiştir. Burada "çubuklu tost" sentezi dil modelinde kontrol edildiğinde toplam tekrar eden "çubuklu tost" sayısının "çubuklu" tekrar sayısına bölümü "çubuklu tost" için gerçekleşme olasılığını vermektedir. Bu sentezin herhangi bir dil modelinde pek az tekrar edeceğini varsayımı yapılırsa, dil modelinin senteze vereceği puan az olacaktır. Devamında ise dil modelinin en iyi alternatif olan "sucuklu" kelimesini önermesi gerekmektedir.

Tablo 1.4: Dil modeli uygulamasına ait basit bir gösterim.

Sentezlenen Cümle	Dil Modeli Uygulanırken
Peynirli ve çubuklu tost sipariş ettim	$P(\text{"çubuklu tost"}) = \frac{Sayı\ (\text{"çubuklu tost"})}{Sayı\ (\text{"çubuklu"})}$
Peynirli ve buçuklu tost sipariş erittim	$P(\text{"sipariş erittim"}) = \frac{Sayı\ (\text{"sipariş erittim"})}{Sayı\ (\text{"sipariş"})}$

Öte yandan dil modeli üretiminde de Türkçe yapısı sebebiyle zorlu bir dil olmaktadır. Sondan eklemeli oluşu kelime üretim kapasitesinin yüksek olmasında ve dil modellerinin aşırı büyümesine sebep olmaktadır.

2. LİTERATÜRDEKİ ÇALIŞMALAR

Konuşma tanıma, önemli ve popüler bir araştırma alanıdır. Günümüzde bu teknoloji insan sesinin payı olan her araştırma ve iş alanında, ihtiyaçları karşılamak ve kullanıcı deneyimini artırmak için kullanılmaktadır. Kaynak yetersizliği bulunmayan hemen hemen her dil için en az bir konuşma tanıma modeli bulunabilmektedir. Akademik çalışmalar ve endüstrinin de katkısıyla bu teknoloji her yıl iyileşmeye ve gelişmeye devam etmektedir. Bununla birlikte modellerin eğitimi için geçmişte ihtiyaç duyulan veri gereksinimleri günümüzde de değişmemiştir. Tam aksine özellikle uçtan-uca yöntemler daha rekabetçi olabilmek için daha çeşitli ve daha büyük veri setlerine ihtiyaç duyabilmektedir [13-15]. Yazıdan görüntü sentezi gerçekleştiren DALL-E 2 eğitiminde 250 milyon görsel yazılı içerikleriyle kullanılmıştır [13]. Bunun yanında büyük dil modellerinden olan ve ChatGPT'nin temelini oluşturan GPT-3, trilyon kelimelik veri setinin filtrelenmesi ile yaklaşım 300 milyar kelime ile eğitilmiştir [14]. Bu miktar öylesine fazladır ki dakikada 238 kelime okuyabilen normal bir insan tarafından tüm metnin tek solukta okunması yaklaşık 2400 yıl sürebilir. Diğer yandan konuşma ve dil tanıma modeli olan Whisper'ın eğitiminde 680 bin saatlik konuşma verisi çevirileri ile kullanılmıştır [15].

2.1. Konuşma Tanımada Güncel Sonuçlar

Otomatik konuşma tanımanın başarısını, sıklıkla bu alanda tercih edilen birkaç veri seti ile değerlendirmek mümkündür. Wall Street Journal (WSJ) ve Librispeech, Tedilium, Aishell1-2 bu veri setlerine örnek olarak gösterebilir [16-20]. Librispeech, İngilizce metinlerin okunması ile derlenen ve okuma konuşması içeren bir veri setidir. Farklı alt setleri olsa ilk duyurulan 1000 saat uzunluğu ile büyük veri setleri arasında kabul edilmektedir [17]. Birçok araştırmacı özgün veri setleri haricinde, çıkarımlarını bu veri seti üzerinden de gerçekleştirerek, geçerli son teknolojiyi geliştirmeye çalışmıştır. Örnek olarak; 80'ler sonrasında konuşma tanıma alanında tercih edilen HMM ve GMM, konuşmacı bağımsız sistemlerin oluşturulması için geliştirilen Speaker Adaptive Training (SAT), yaygın başarısı ile birçok alanda popülerlik kazanmış Deep Neural Network (DNN) ve zamanla değişen girişlerde iyi performans gösteren LSTM-RNN için farklı zamanlarda sonuç kaydetmiş çalışmalar bulunmaktadır [17, 21, 22].

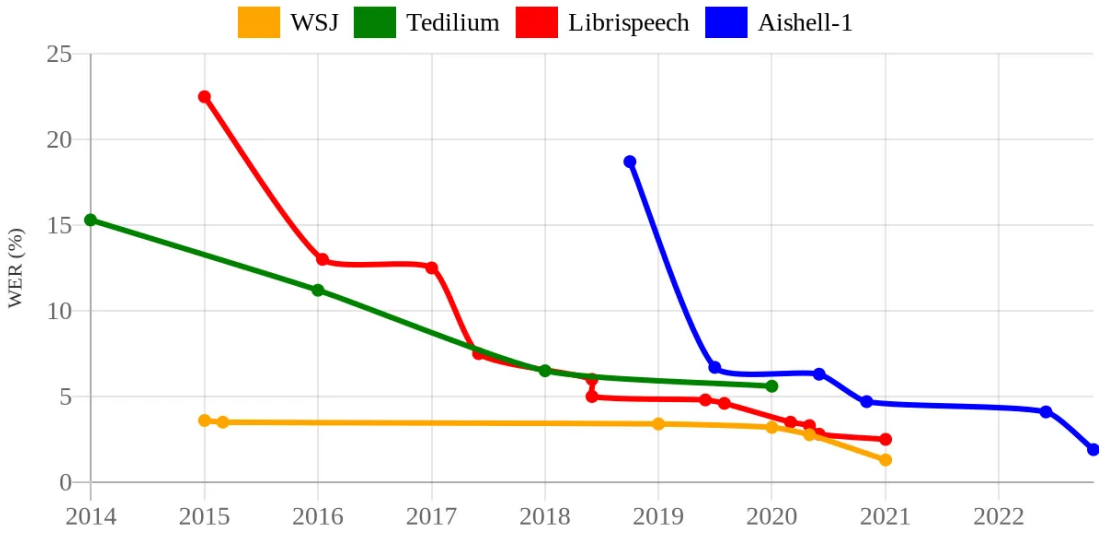
Bunun yanında, CTC sayesinde farklı modellerin bir arada eğitilmesiyle uçtan-uca sistemlerde Transformer ve Conformer yapıları gündeme gelmiştir. Bu yöntemler geçerli son teknolojiyi ileriye götüren gelişmelere aracılık etmiştir [22, 23]. Karita ve arkadaşları yazılımlarına Transformer yapısını dahil ederek RNN'lere karşı daha başarılı sonuçlar alındığını göstermiştir. Bu çalışmada özellikle Librispeech 1000 saatlik corpus özelinde geliştirilen model, Kaldi modelini %1,7 farkla geçerek %97,4 kelime hata oranına erişmiştir [22]. Gulati ve arkadaşları ayrıca bir yıl sonra Guo ve ark. Conformer eklemesi ile yaparak kelime hata oranını sırayla %2,0 ve %1,9'a indirmişlerdir [23, 24]. Tablo 2.1'de Librispeech test seti için eski ve yeni sonuçlar ile listelenmiştir.

Tablo 2.1*:* Librispeech test kümesi için rapor edilen WER değerleri.

Kaynak	Metot	WER
Lüscher ve ark. [21]	GMM	24.1
Lüscher ve ark. [21]	GMM + SAT	8.04
Vassil ve ark. [17]	DNN	5.51
Lüscher ve ark. [21]	LSTM + SAT	4.4
Karita ve ark. [22]	RNN	3.3
Radford ve ark. [15]	Transformer	2.7
Karita ve ark. [22]	Transformer	2.6
Gulati ve ark. [23]	Conformer	2.0
Guo ve ark. [23]	Conformer	1.9

2.2. Konuşma Tanımada Veri Setlerinin Önemi

Küresel veri tabanları görüntü ve video gibi evrensel dile sahip veri gruplarında yeterli olsa da dile ve bölgeye bağlı; yazı ve ses kaydı gibi veri gruplarındaki eksiklikler nispeten fazla hissedilebilir. Bu nedenle araştırmacıların hızlarını artıracak bu gruplardaki kaynaklar büyük değer taşımaktadır. Şekil 2.1'de ise Librispeech yanında WSJ, Tedilium ve Aishell veri setlerini kullanan çalışmaların ortaya koyduğu WER değerleri yıllara göre verilmiştir. Yayınlanma zamanından başlayarak bir referans olan bu veri setleri ve ilgili çalışmaları; araştırmacılara ortak test alanı sağladığı gibi, literatürün kümülatif bilgi birikimine de katkı sağlamaktadır. Şekil 2.1'deki hata oranlarından, çalışmalar sonunda üretilen modellerin yıllara göre daha yüksek başarı ortaya koyduğu gözlemlenebilir. Farklı veri setleri dillere dair problemleri göz önüne koyarken, araştırmacılara daha farklı yaklaşımları test etme imkânı vermiştir.



Şekil 2.1: Yıllara göre bazı veri setleri için rapor edilen WER değerleri [25, 26].

2.3. Türkçe Konuşma Veri Setleri

Dile bağımlı çalışmaların sonucunda sunulan bazı açık veri setlerine Open Speech and Language Resources (OpenSLR) üzerinden ulaşılabilmektedir [27]. OpenSLR birçok dil için metin, ses, konuşma veya yazılım materyali barındırır. Tablo 2.2'de OpenSLR'a ait Almanca, Çince, Fransızca, İngilizce, İspanyolca, Rusça ve Türkçe konuşma verilerinin uzunluklarına göre dağılımı derlenmiştir. Bu derleme yapılırken; aynı dile ait farklı lehçe ve ağızlar birleştirilmiş, aynı veri setine ait son ve çeviri ile doğrulanmış versiyonlar esas alınmıştır. Bunun yanında, Text-to-Speech (TTS) ya da Speech-to-Text (STT) kategori ayrımı yapılmamış ve ses kalitesi göz önüne alınmamıştır.

Tablo 2.2: Dillere göre toplam konuşma verisi uzunlukları - OpenSLR.

Saat	Dil
14730	Çince
2056	İngilizce
1338	Rusça
83	İspanyolca
26	Almanca
10	Fransızca
10	Türkçe

Başka bir kaynak olan Mozilla Foundation tarafından ücretsiz olarak yürütülen ve sunulan Common Voice projesi de araştırmacılara veri yetersizliğinde destek olmaktadır [28]. Common Voice, 2022 itibariyle 100'den fazla dile ve 23 bin saatten fazla ses kaydına ulaşmış günümüzdeki en büyük kitle kaynaklı, doğrulanmış, açık veri kümesi konumundadır [29]. Tablo 2.3'te bu veri kümesine ait 7 farklı dildeki doğrulanmış ses kaydı uzunlukları verilmiştir.

Tablo 2.3: Dillere göre toplam konuşma verisi uzunlukları – Common Voice

Saat	Dil
3286	İngilizce
2279	İspanyolca
1366	Almanca
1221	Çince
1077	Fransızca
257	Rusça
111	Türkçe

Açık kaynaklardan elde edilebilecek Türkçe konuşma kayıtlarının yanında üniversiteler arası bir konsorsiyum olan Linguistic Data Consortium (LDC) ise araştırmacıların verilere daha kategorize edilmiş ve geniş ölçekte ulaşabilmelerini sağlamaktadır. Fakat maalesef LDC üyeliği kâr amacı gütmeyen üniversite ve diğer organizasyonlar için dahi yüksek bir bedele sahiptir, bu bedel standart üyelik için 2400$ ve abonelik için 3850$ olarak görünmektedir [30]. Tablo 2.4'te LDC, Common Voice ve OpenSLR üzerinden erişilebilenlerin yanında ve bazı araştırmacılara ait çevirileri yapılmış Türkçe konuşma veri setleri verilmiştir.

Sonuç olarak Türkçe konuşma kayıtlarına ve çevirilerine tümüyle dahi ulaşabilen araştırmacılar son teknoloji modeller üzerinde çalışmak istediklerinde İngilizce, Rusça veya Çince gibi dillerin aksine veri yetersizliğiyle karşılaşabilmektedir. Tablo 2.2 ve 2.3'te Türkçe için bu fark görülebilirken, Tablo 2.4 eklendiğinde dahi rekabetçi bir veri boyutuna ulaşılamamaktadır. Buradan da yola çıkarak; daha rekabetçi Türkçe konuşma tanıma, daha kaliteli ses sentezi ve ses çevrimi gibi modellerin gelişimi için daha fazla doğrulanmış veriye ihtiyaç olduğu açıktır.

Tablo 2.4: Türkçe konuşma tanımada kullanılmış bazı veri setleri.

Saat	Çalışma	Erişim
105	Common Voice Corpus 14.0 - Turkish [28]	Common Voice
5,6	Middle East Technical University Turkish Microphone Speech v1.0 – Turkish [31]	LDC
130	Turkish Broadcast News Speech and Transcripts – Turkish [32]	LDC
9,3	2009 NIST Language Recognition Evaluation Test Set - Multilangual [33]	LDC
213	IARPA Babel Turkish Language Pack IARPA-babel105b-v0.5 – Turkish [34]	LDC
18.6	Multi-Language Conversational Telephone Speech 2011 - Turkish [35]	LDC
8,5	2011 NIST Language Recognition Evaluation Test Set [36]	LDC
6,1	A New Database For Turkish Speech Recognition On Mobile Devices And Initial Speech Recognition Results Using The Database – Turkish [37]	-
39,2	Automatic Speech Recognition System Adaptation For Spoken Lecture Processing - Multilangual [38]	-
350	Web Service-Based Turkish Automatic Speech Recognition Platform - Turkish [39]	-
10	MediaSpeech: Multilanguage ASR Benchmark and Dataset – Multilangual [40]	OpenSLR
895,3		

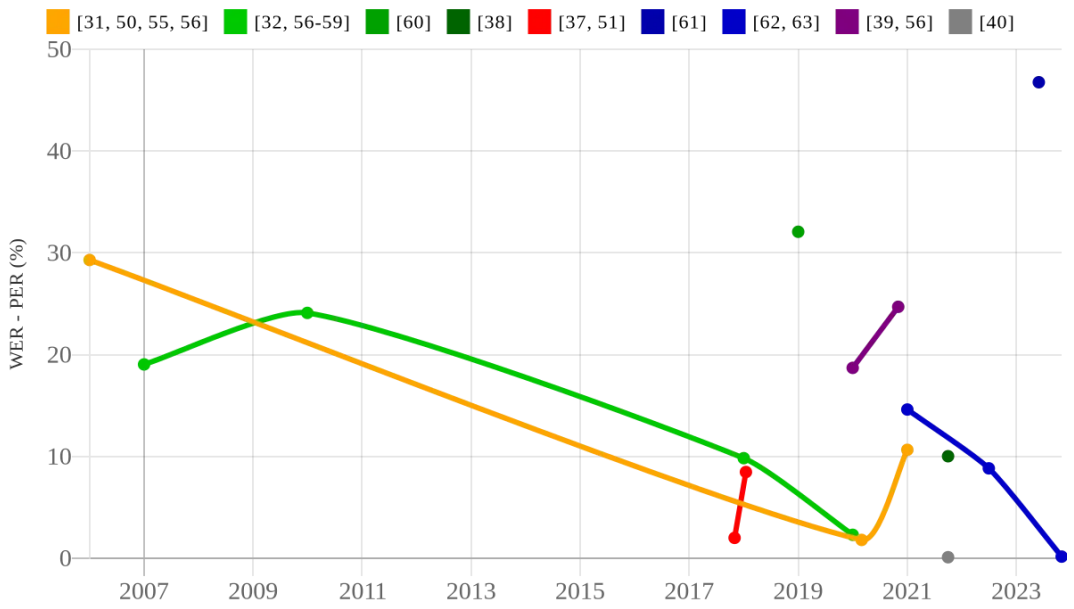
Bazı araştırmacılar açık veri kaynaklarını filtrelemiş, yabancı ve yerli literatüre katkı sağlamıştır. Zhang, Lv ve ark. [41] çalışmasında Youtube üzerindeki Çince videoları ve podcastleri kullanarak 10000 saat çevirisi yapılmış, 2400+ saat kalitesiz-çevirili ve 10000 saat çevirisi yapılmamış olarak toplamda 22400+ saatlik bir derlem oluşturmuşlardır. Bu derlemde tercih edilen video ve podcastler farklı konuşma türleri, gürültü ortamları, konu ve alanları bir araya getirmek üzere özellikle seçilmiştir. Bu çalışma 10 farklı veri kategorisini bir araya getirerek, yüksek çeşitlilik ve gerçek dünya verisi sağlayarak, bilinen en büyük ölçekli Mandarin derlemleri olan 520 saatlik Aishell-1 ve 1000 saatlik Aishell-2'yi geride bırakmıştır [18-19]. Çalışmada kayıtların yüksek doğruluklu çevirilerinin üretiminde öncelikli olarak endüstriyel başarısını kanıtlamış bir konuşma tanımı kullanılırken, ikincil doğrulama katmanı olarak video altyazılarının Optical Character Recognition (OCR) okunması yöntemi tercih edilmiştir. Diğer yandan Kolobov ve ark. [40] çalışmasıyla yine Youtube üzerinde bazı haber kaynaklarından Fransızca, Arapça, İspanyolca ve Türkçe için her biri 10'ar saatlik derlemler oluşturmuştur. Bu veri seti Türkçe için Türk araştırmacıların ortaya koymadığı erişime açık bir kaynak olarak ilgili alanda literatüre katkı sağlamaktadır. Türkçe kayıtlar için Fox Haber ve Show Ana

Haber tercih edilmiştir [40]. Bu sebeple veri seti daha çok haber konularını içeren, temiz ve [41]'ye kıyasla nispeten çeşitliliği daha az kayıtlar sağlamaktadır. Benzeri şekilde bu çalışma için de yazarlar bazı hazır konuşma tanıma modelleri ile ön çevirileri sağlamışlardır. Fakat farklı olarak çevirileri bir takım gerçekleştirmiştir. Common Voice doğrulanmış Türkçe segmentini çeviricileri kontrol için kullanarak farklı bir doğrulama yöntemi elde etmişlerdir [40].

2.4. Geçmişten Bugüne Türkçe Konuşma Tanıma

Türkçe konuşma tanıma için yapılan araştırmalar göz önüne alındığında literatürdeki yöntemlerin birçoğunun araştırıldığını fakat henüz çok yeni teknolojiler olması ve hali hazırda büyük miktarda Türkçe konuşma veri setine ulaşılamamasının etkisiyle Transformer ve Conformer için fazla çalışma olmadığını söylememiz mümkün olacaktır. Literatüre bakıldığında Türkçe konuşma tanıma için yapılan araştırmalarda Türkçe'ye özel yapısal zorluklara ve bunların aşılmasına odaklanıldığını görülmektedir [42-48]. Son 10 yılda ise bu alanda gelişen son teknoloji ile nispeten daha yenilikçi yaklaşımlar ile deneyler sürdürülmüştür [39, 49-54].

Türkçe konuşma tanıma üzerine yapılan araştırmaları ve aralarındaki bağlantıları görebilmek için hazırlanan literatür özeti yıllara ve veri setlerine göre gruplanmıştır. Şekil 2.2'de bu ilişkili çalışmaları aynı renk ile gösterilmiştir.



Şekil 2.2: Veri seti yönünden bağlantılı Türkçe konuşma tanıma model başarıları ve çalışmaları.

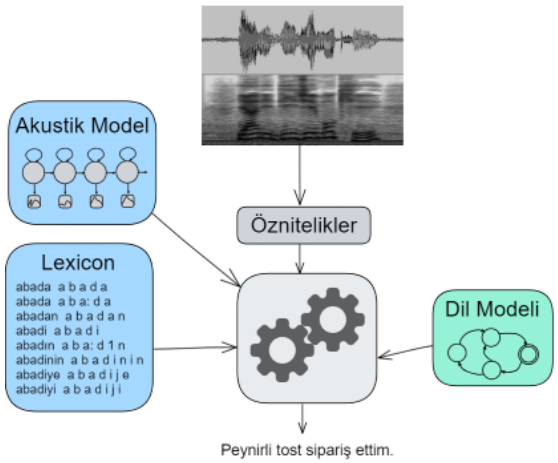
Şekil 2.2 için sonuçlar toplanırken model eğitimlerinin aynı veri seti üzerinde yapılmasına dikkat edilmemiştir. Çalışmalar farklı zamanlarda, farklı yaklaşımları içerdiğinden ve değerlendirme metrikleri, veri seti kullanım şekilleri farklı olabileceğinden, bağlantılardaki sürekli iyiye gidiş gözlemlenmeyebilir. Benzeri renk tonlarında olan fakat ayrıklık oluşturan çalışmalar yeni yaklaşımları denedikleri için bu şekilde gösterilmiştir.

Türkçe için yapılan çalışmalarda 2000'lerden başlayarak Sphinx [64], HTK [65] sonrasında ise Kaldi [66] kullanılarak birçok model geliştirilmiştir. Bu modellerde 2011'e kadar HMM-GMM mimarisi kullanılırken sonrasında HMM-DNN yapıları da tercih edildi [51] . Alternatif olarak araştırmacılar kendi yazılımlarını geliştirdiler. Palaz ve arkadaşları TREN isimli konuşma tanıma yazılımını 2005'te yayınladılar [67]. Bu yazılım MFCC öznitelikleri ile HMM akustik modellemesini taban almaktaydı. Sonraki yıllarda Türkçe için dile bağlı bazı yapısal zorlukların çözümü için birçok araştırma gerçekleştirildi. Türkçenin morfolojik yapısı sondan eklemeli oluşu sebebiyle karmaşık olduğu için konuşma tanıma sistemlerinde inanılmaz boyutta kelime olasılığının başarılı şekilde çözülmesi beklenmiştir [44, 46]. Bu durum çözülmesi güç bir problem yaratmıştır. 2002'de Dutaağacı çalışmasında bu problemi konu olarak yüksek lisans tezini tamamladı [68]. 2005'te Büyük yine aynı konuyu ele alarak alt kelimeler oluşturmak vasıtasıyla problemin çözümünü doktora tezinde ele almıştır [46]. 2009'da Arısoy konu ilgili dil modellerinin geliştirmesi üzerine yüksek lisans tezini vermiştir [48]. Literatüre bakıldığında 2000-2014 yılları arası Türkçe için bu problemin çözümü birçok kez ele alındı [42-48]. 2018'de Asefisaray doktora teziyle uçtan-uca yöntemlerle geleneksel yöntemleri kıyaslayarak ve ortaya Türkçe uçtan-uca modeller koyarak Türkçe konuşma tanımasının ilerleyişine katkı sağladı [54]. Kimanuka ve Büyük çalışmalarında DNN-HMM ile GMM-HMM mimarisini Kaldi üzerinde test ederek kıyaslamıştır [51]. Tombaloğlu ve Erdem Türkçe için LSTM ve Gated Recurrent Unit (GRU) tabanlı sistemleri ele almışlardır [50]. Oyucu ve arkadaşları [39]'de web servis tabanlı Türkçe konuşma tanıma sistemini geliştirmişlerdir. Bu sistem MFCC öznitelikleri ile GMM-HMM tabanında konuşma tanıma hedeflemiştir. Bunların yanında Arslan ve Barışçı [49]'de Türkçe için konuşma tanıma odağında geniş bir araştırma gerçekleştirmişlerdir. Dikici ise çalışmasında [69] gözetimli, gözetimsiz ve yarı gözetimli öğrenme yöntemlerini kullanıp dil modeli entegrasyonu ile konuşma tanıma üzerinde çalışmıştır. Ahmed [70]'de Çağrı merkezlerine ayrıca konuşma tanıma gerçekleştirilmiştir.

Çalışmasında olasılıklar gücüyle bilinen HMM kullanılmıştır. Arslan çalışmasında [71] hatalı ASR çıktılarını LSTM-GRU ile düzenlemeye yönelik çalışmıştır. Kutucu [72]'de CNN kullanımı ile Türkçe ASR oluşturmuştur. Öte yandan Oyucu [73]'da geleneksel yöntemleri kullanan ve mobil cihazlarda verimli şekilde çalışabilen bir Türkçe konuşma tanıma sistemi geliştirmiştir. Bu çalışmasında Kaldi'den destek almıştır. Tombaloğlu doktora tezinde [74] Kaldi ile LSTM-GRU modellerini kullanarak konuşmayı işaret diline çeviren bir sistem oluşturmuştur. Fakhan önemli bir çalışma gerçekleştirmiş ve güncel Transformer metotlarını ve bazı farklı modelleri kullanılmasıyla farklı veri gruplarından öğrenilen bilgiyi Türkçe ASR üretimi amacıyla kullanarak performansı artırmıştır [38].

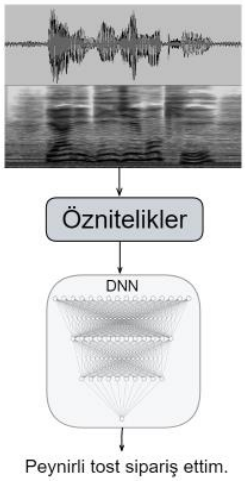
2.5. Konuşma Tanımada Geleneksel ve Yenilikçi Yöntemler

Konuşma tanıma için geleneksel yaklaşım dönemini, diziden diziye (seq2seq) ve end-to-end yöntemlerin popüler hale geldiği noktadan geriye HMM-GMM yapılarının kullanılmaya başlandığı dönemlere kadar olan yakın geçmiş ile sınırlandırmak mümkündür. Bu dönemde veri seti büyüklüğündeki azlık, hesaplama güçlükleri ve daha önemlisi yenilikçi yaklaşımlara dair eğitim güçlükleri son teknolojinin bir adım öteye gitmesinde engel oluşturmuştur [5]. Geleneksel yöntemleri HMM-GMM, sonrasında HMM-DNN yapıları ve daha sonrasında ise klasik DNN yerine koyulan CNN ve hatta RNN çalışmaları olarak hatırlamak mümkündür [5, 75]. Bu tip modeller için dile özgü birçok parametre modellerin başarısına etki etmektedir. Bu durum Şekil 2.3'te gösterilmiştir. Şekildeki Lexicon, Akustik model ve Dil modeli, ilgili dil için dil için çoğu zaman ekstra çalışmaları gerektirmektedir.



Şekil 2.3: Geleneksel konuşma tanıma yöntemleri için bir blok diyagram.

Öte yandan yenilikçi yöntemlerde, Şekil 2.3'teki alt modellerin kullanımı büyük oranla azalmıştır. Bazı yöntemler tamamen ses girdisini dizi halinde ele alıp kelime dizisi çıktıları üretirken bazı yöntemler ise çıktı üzerinde dil modeli çalıştırarak veya giriş özniteliklerinde fonemlere yer vererek başarıyı artırmayı hedeflemiştir [7, 15, 75]. Yenilikçi yaklaşım birçok farklı şekilde olabilmektedir, temsilen genel yapı Şekil 2.4'te gösterilmektedir.



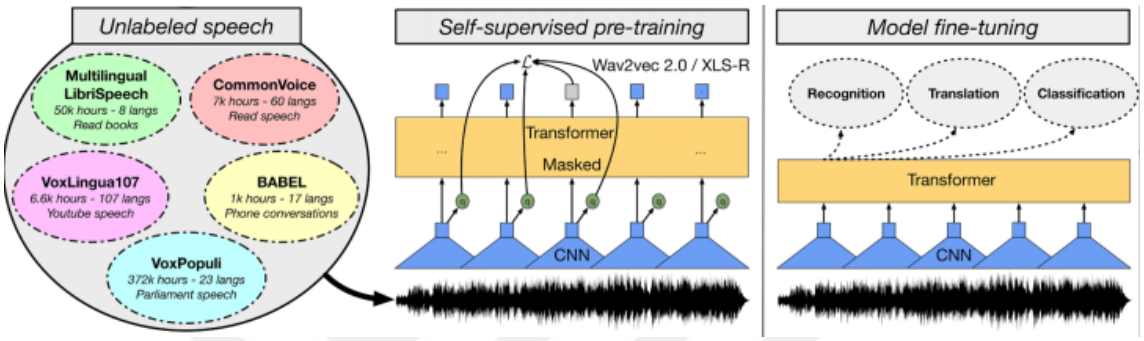
Şekil 2.4: Yenilikçi konuşma tanıma yöntemleri için basit bir temsil.

2.6. Konuşma Tanımada Gözetimli ve Gözetimsiz Öğrenme

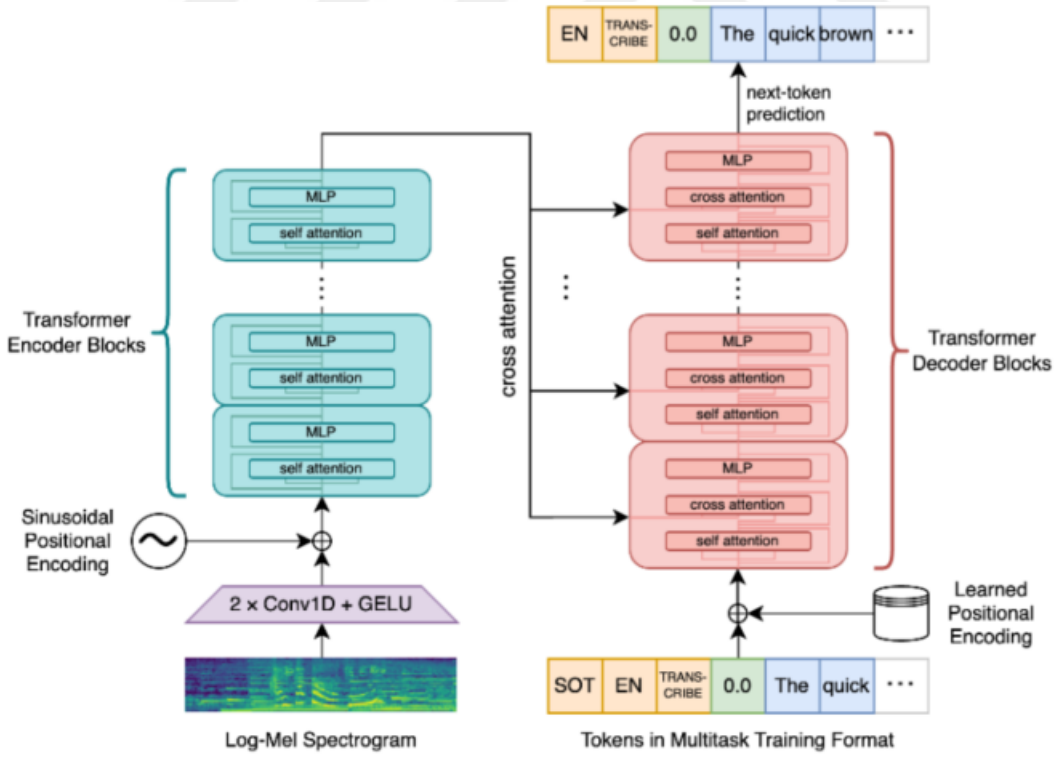
Gözetimli öğrenme, etiketli veri setlerinin kullanımına dayanırken, gözetimsiz öğrenme etiketsiz verilerin kullanımına dayanmaktadır. Çoğunlukla gözetimsiz öğrenmede modeller daha sonradan etiketli veriler ile ince ayar yapılarak son haline getirilmektedir. Günümüzde geleneksel veya uçtan-uca modeller için de geçerli olmak üzere, konuşma tanıma için son teknoloji modeller halen gözetimli öğrenme yolunu izleyen modeller olmakta veya bu yaklaşım daha fazla umut vadetmektedir.

İki yöntemi son yıllarda artan gözetimsiz öğrenme temelli çalışmalara dayanarak kıyaslamak mümkün hale gelmiştir. Güncel bir örnek olarak; gözetimli öğrenme noktasında 680 bin saatlik çok dilli, etiketli veri kullanan Whisper karşısında gözetimsiz öğrenme ile 436 bin saatlik çok dilli, etiketsiz veri kullanan Wav2vec2 mimarisini taban alan XLS-R kıyaslandığında Librispeech temiz test kümesinde WER değerleri sırasıyla 2.7 ve 5.9 olmaktadır [15, 76]. Bunun yanında ise orijinal Wav2vec2 çalışmasında ise

Librispeech veri seti ile yapılan eğitimde aynı test kümesi kullanıldığında rapor edilen WER değeri Whisper ile aynıdır [15, 77]. Bu noktada Whisper'ın farkı ise farklı test koşullarında açıkça görülmektedir, örneğin; farklı veri kümeleri üzerinden yapılan kıyaslamada Whisper, Wav2vec2 modeline göre %55 daha yüksek başarı gösterirken, farklı gürültü seviyeleri altında da ücretli veya ücretsiz olarak sunulan birçok modele göre daha yavaş performans düşüşü göstermektedir [15, 78-81].



Şekil 2.5: Wav2vec2 Mimarisine ait blok diyagram [76].



Şekil 2.6: Whisper'ın çoklu dil ve görevli sonuç üreten mimarisine ait blok diyagram [15].

3. MATERYAL VE YÖNTEM

3.1. TURKSPEECH Veri Setinin Hazırlanması

Veri setinin hazırlanmasında olabildiğince farklı konuşmacılara ve ses arka planına erişilmesi planlanmıştır. Bunun için ne sadece profesyonel ortamlardan konuşma veya okuma kayıtları ne de yalnızca gerçek dünya verisi seçilmiştir. Yararlanılan ses kaynakları; meclis konuşmaları, haber bültenleri, podcastler ile sohbetler ve monologlar, maç yorumları, online oynanan oyunlara ait yorumlar, kitap okumaları ve toplantı görüşmeleridir. Kaynaklara ait bağlantılar tümüyle Youtube üzerinden kaydedilmiştir. Toplamda 24 kategoride 1177 kaynaktan yararlanılmıştır. Bu kategoriler ve barındırdıkları kaynak sayıları Tablo 3.1'de gösterilmektedir. Her kategori için toplamlar kaynak sayılarını gösterir, bunun dosya sayısıyla karıştırılmaması gerekir. Ek olarak bir kaynak yıllar sürebildiği gibi saatler sonunda sonlanmış da olabilir.

Tablo 3.1: Veri seti derlemesi: Kaynak kategorileri ve toplam sayıları.

Kategori	**#**	**Kategori**	**#**
Aile	5	Güncel	59
Astroloji	6	Haber	112
Basın	7	İş	87
Bilim	122	Müzik	34
Edebiyat	62	Oyun	7
Eğitim	45	Psikoloji	26
Eğlence	62	Radyo	18
Ekonomi	19	Röportaj	6
Felsefe	13	Sanat	117
Gastronomi	6	Sinema	4
Gelişim	61	Sohbet	191
Girişimcilik	16	Spor	92

3.1.1. Ses Kayıtlarının İndirilmesi

Ses kaynakların kategorize edilmesi ve bağlantıların kaydedilmesi sonrası kayıtlar Python aracılığıyla wget yazılımı kullanılarak indirilmiştir. Bazı kayıtlar maalesef geçersiz bağlantıya sahip olduğundan boş olarak kaydedilmiştir, bu kayıtlar daha önceden silinmiştir. Dosya sayısı yüksek ve her dosya en az 16 kHz kaliteli ses barındırdığından

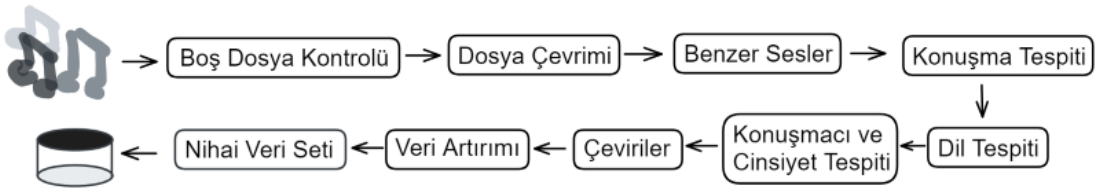
kayıt depolaması noktasında 12 TB alana sahip performans diski tercih edilmiştir. Bunun yanında ise indirilen bu kayıtlarını işlerken 2 TB büyüklüğünden performans diskleri tercih edilmiştir. Kaynak çeşitliliği ve çıktı formatları Tablo 3.2'de gözlemlenebilir. Kayıtların çıktı frekansını 16 kHz gibi optimum bir değerde tutarak kaynak israfından kaçınırken ses kalitesinden ödün verilmemiştir.

Tablo 3.2: Orijinal veri ilk formatı ve son formatı.

	İlk Format	**Son Format**
Uzantı	.mp3 .m4a .mp4	.wav
Frekans (kHz)	48, 44.1, 32, 22.5	16
Kodek	AAC, MPEG	PCM S16LE
Kanal	2, 1	1
Sıkıştırma	✓	×

3.1.2. Ön İşleme Adımları

Derlemeyi de içerek şekilde takip ettiğimiz ön işleme adımları Şekil 3.1'de gösterilmiştir.



Şekil 3.1: Veri seti ön işleme adımları.

3.1.3. Benzer Seslerin Elenmesi

Benzer veya aynı kayıtların nihai veri setinden uzaklaştırılması hem çeşitliliği artırması hem de aynı verilerin tekrar tekrar öğrenilmesini engelleyebilir. Ayrıca birden fazla kopya kaynak israfına da yol açabilir. Bu adımı tam anlamıyla geçebilmek için indirilen dosyaların detaylı şekilde analizi gerçekleştirilmiştir. Bu aşama başarılı olsa dahi dosya içeriğindeki uzun süreli benzerlikler problem yaratabilir. Bu nedenle yapısal ve içeriksel olarak iki tür yaklaşım tercih edilmiştir. İki yöntem ile toplamda 4138 kaydın veri setinden temizlenmesi sağlanmıştır.

3.1.3.1. Yapısal Analiz

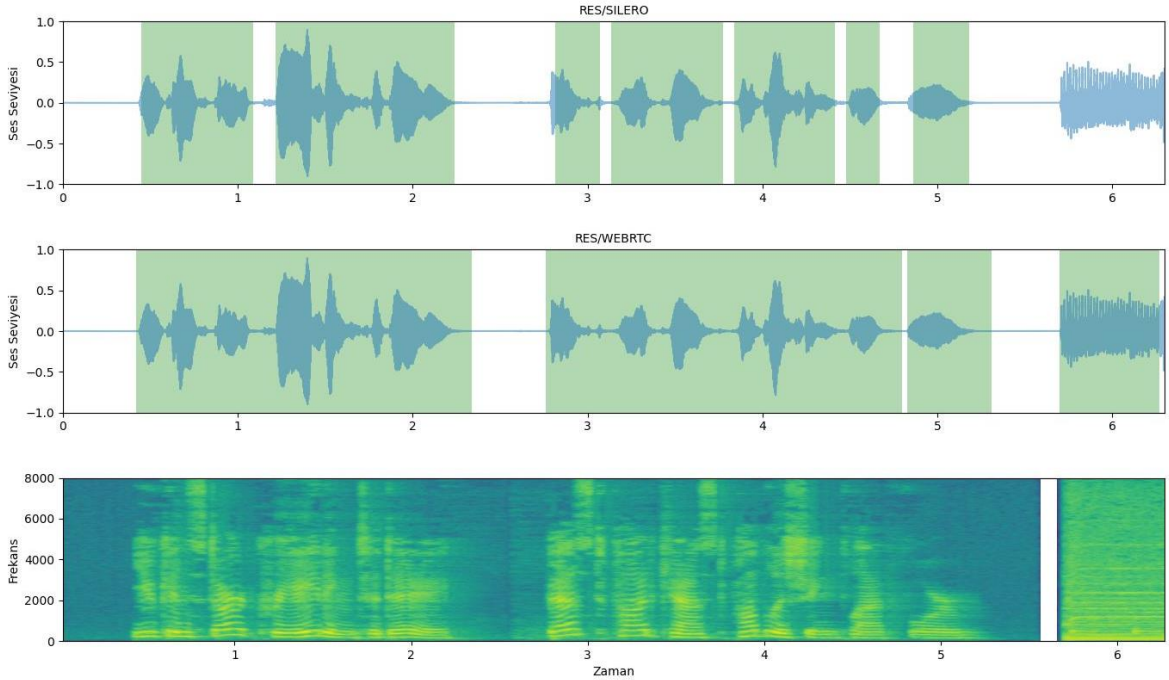
Bu kısımda benzeri dosya isimleri kontrolünden sonrası dosyaların boyut, bit dizilimi ve kayıt uzunlukları gibi bütünüyle görünümleri ele alınmıştır. Boyut ve kayıt uzunluğuna göre sıralanan verilerden birbirine fazla yakın olanların bit dizilimleri kontrol edilerek verinin tekrarlardan uzaklaştırılması sağlanmıştır. Fakat bu yöntem aynı ses kaydının bit dizilimi farklı olsa 16 kHz veya 44.1 kHz gibi varyantlarını tespit edemeyebilir.

3.1.3.2. İçeriksel Analiz

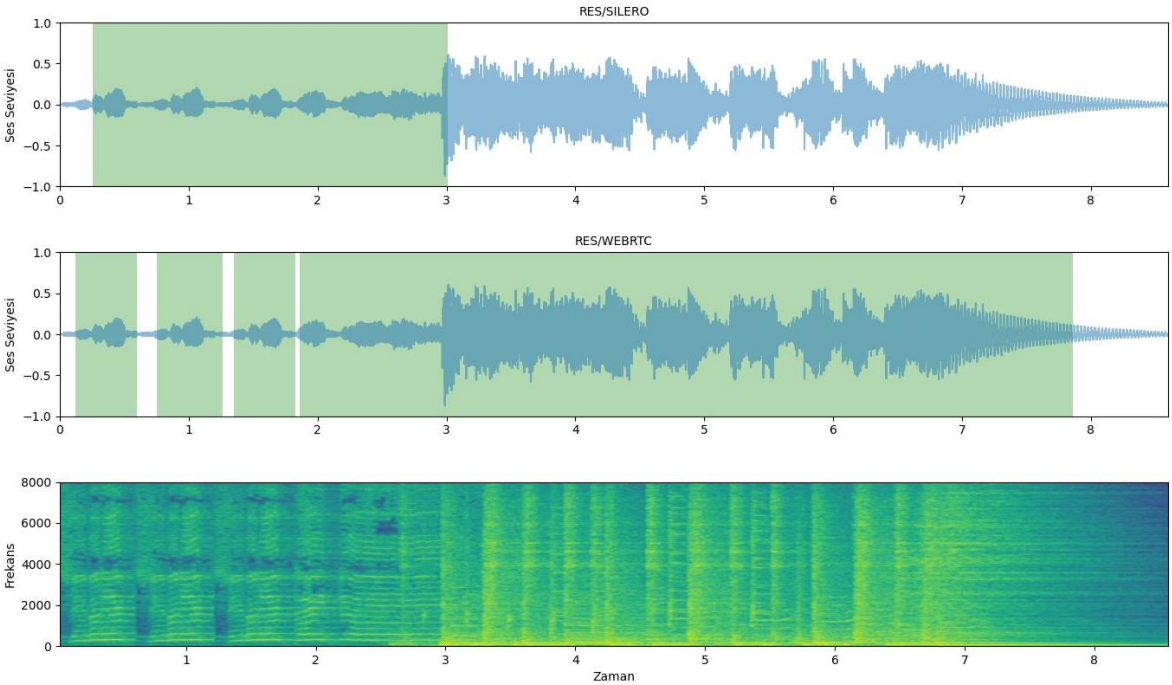
Tekrarlı kayıtlardan yapısal aşamasından kurtulanları bu analiz yöntemi ile tekrardan tespit edilmeye çalışılmıştır. Bu tür bir karşılaştırma ses dosyalarının ancak frekans uzayında yapılabildiğinden MFCC [12] özniteliklerini elde ettiğimiz şüpheli kayıtların karşılaştırmaları gerçekleştirilmiştir.

3.1.4. Ön İşleme Adımları Konuşma Sesi Tespiti

Yüksek miktarda ses verisini işlemek pratik bir yaklaşımda doğru kabul edilmeyebilir. Konuşma veri setlerinde konuşma içeren kısımlar verinin kalanına göre daha önemli olabilir. Arka plan seslerini mümkün olduğunca azaltılması, eğitilecek modellere daha amaca yönelik verinin sağlanması ve kaynak kullanımının yönetilmesi amacıyla kayıtlara Ses Aktivite Tespiti (VAD) uygulanmıştır. Performanslı ve etkisini günümüzde tüm web sayfalarındaki medya akışlarında gösteren WebRTC örnek bir VAD barındırmaktadır [82]. Daha çok optimum ve performansa yönelik çalışan bu yazılım bazı gürültü durumlarında yeterince başarılı olamayabilir. Bu noktada daha güncel ve derin öğrenme arka planına sahip olan Silero VAD [83] modelinin kullanılması uygun bulunmuştur. Silero temelde bir VAD olmasına rağmen eğitiminde tecrübe ettiği farklı senaryolar sayesinde başarılı bir konuşma aktivitesi dedektörü olabilmektedir. Ses yerine konuşmayı tespit etmeye yönelmek, burada üzerinde çalışılan veri seti için oldukça kritiktir. Bunun sebebi ise kaynak verilerimizin yoğun şekilde müzik, ortam sesleri veya teknik nedenli gürültülere sahip olmasıdır. Şekil 3.2 ve 3.3'te Silero ve WebRTC VAD karşılaştırması bu gürültülere sahip bir veri seti örneği üzerinden yapılmıştır. Veri setindeki tüm konuşma aktiviteleri en kısası 3-5 saniye ve en uzunu 20-60 saniye olacak şekilde orijinal kayıtlardan kesilerek tekrar kaydedilmiştir. Bu aşamada yaklaşık 107503 dosya için 1 milyon 750 binden fazla konuşma parçacığı elde edilmiştir.



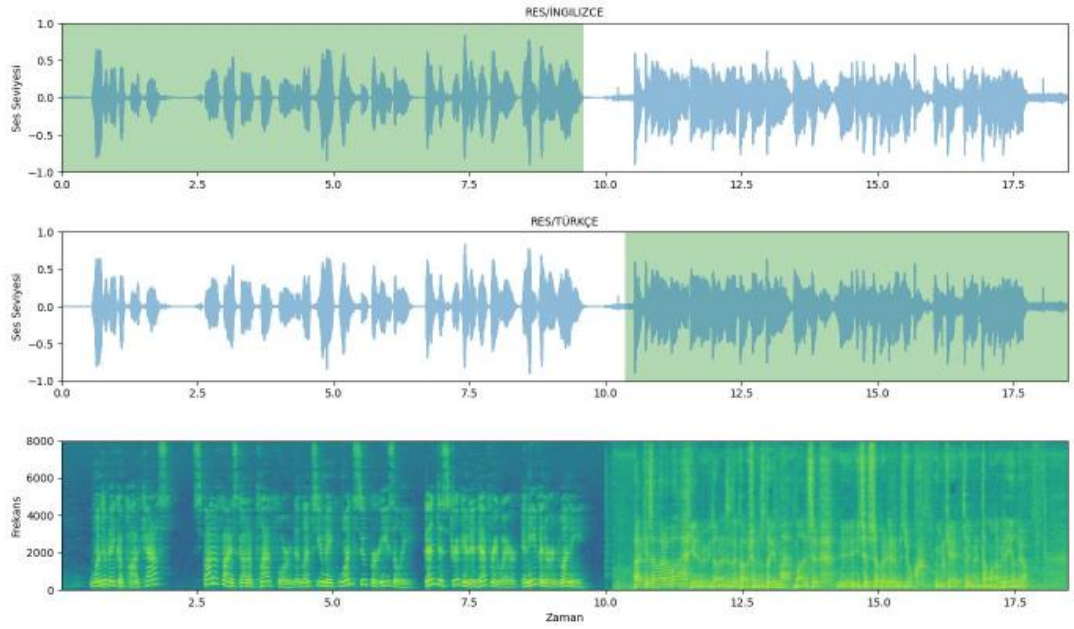
Şekil 3.2: Gürültü açısından Silero ve WebRTC karşılaştırması.



Şekil 3.3: Müzikal arka plan açısından Silero ve WebRTC karşılaştırması.

3.1.5. Dil Tespiti

Konuşma tespiti sonrası insan sesine odaklanarak gerçekleştirilen filtrelemede kayıt içerisinde varsa farklı dillere ait kısımlar ise bu aşamanın oluşmasına neden olmuştur. Her ne kadar kelimeler özelinde bir filtreleme yapmak yanlış olsa da uzun süreye sahip yabancı içeriklerin veri setinden atılması gerekmektedir. Dil tespiti için 680 bin saat farklı veri grupları ile eğitilmiş olan Whisper'ı [15] ve 95 farklı dili destekleyen Silero dil sınıflandırıcısı [83] tercih edilmiştir. Şekil 3.4'te örnek bir kaydın ilk dakikalarında geçen yabancı konuşmalar Whisper'ın desteği sayesinde işaretlenerek gösterilmiştir

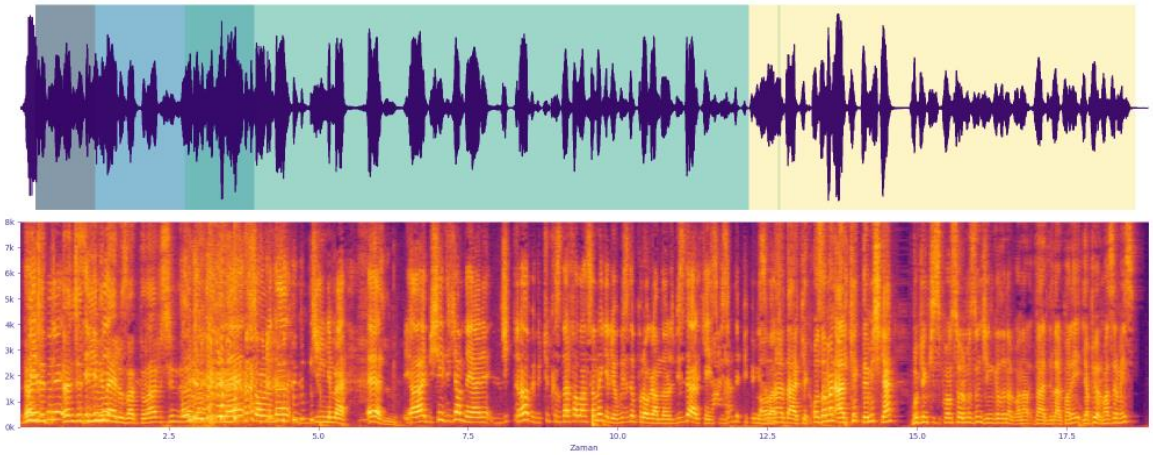


Şekil 3.4: Dil tespitinde elde edilen bir sonuç, kaydın ilk kısmı İngilizce iken devamı Türkçe.

3.1.6. Örtüşme Kontrolleri

Daha önce de belirttiğimiz gibi özellikle monolog olmayan konuşma kayıtları ses örtüşmelerine öncülük etmektedir. İlgili kayıtlar gerçek dünya verilerine yakın olduğundan yüksek önem derecesine de sahiptir. Bu amaçla veri setinin hem rekabetçi kalabilmesi hem de aşırı örtüşmüş ses kayıtlarının işaretlenmesinde pyAnnote [84, 85] kullandık. pyAnnote bir konuşmacı etiketleme yazılımı olarak Bredin ve arkadaşları tarafından oluşturulmuştur. Bu model, anlık olarak 4 kadar konuşmacının ayrımını yapabilir. Şekil 3.5'te veri setimizden örnek bir kayda ait pyAnnote sonuçları gösterilmiştir. Bu kayıtta 4 farklı konuşmacının yüksek oranda konuşma örtüşmesi

bulunmaktadır. Her ne kadar Şekil 3.5'teki segmentasyonda konuşma aktiviteleri ayrı görünüyor olsa bu ses kaydındaki örtüşmelerin fazla olduğu gözlemlenmiştir. Bu sebeple segmentasyonların birbirlerine bu kadar yakın olduğu ses kayıtları daha sonradan veri setinin zorlu bir alt kümesi olarak tekrar test edilmek üzere ilk versiyondan ayrı tutulmuştur.



Şekil 3.5: pyAnnote ile segmentasyon sonuçlarının gösterimi. Üstte dalga formu ve konuşmacılar, altta ise spektrogram.

3.1.7. Cinsiyet ve Diğer Metrikler

Konuşma veri setleri barındırdıkları zengin veri çeşitliği nedeniyle farklı amaç taşıyan modelleri besleyebilir veya test kümesi oluşturabilirler. Örnek olarak ses parçalarından cinsiyet ve yaş tespiti, ses kalitesinde iyileştirme veya gürültüden arındırma, aynı konuşmacılara ait seslerin ses klonlaması veya ses sentezinde kullanılması. Veri setinin bunun gibi çalışmalarda kullanılmasını desteklemek üzere bir metadata havuzu oluşturulmuştur. Konuşmacıların ilgili parça içerisinde tespiti ve etiketlenmesi için pyAnnote yeniden kullanılmıştır. Konuşmacı cinsiyetlerinin eldesinde [86]'da önerilen 4690 adet etiketli veri ile eğitilmiş model kullanılmıştır. Dosyalara ait özellikler için frekans, uzunluk, format gibi metrikler kullanılmıştır. Bunların yanı sıra ses kayıtları için daha detayları bilgilerin sağlanması adına [86] çalışmasındaki veri madenciliği için kullanılan metrikler kullanılmıştır. Son olarak ise yeni bir çalışma olan ve Wideband Perceptual Estimation of Speech Quality (PESQ) [87], Short-Time Objective Intelligibility (STOI) [88] ve Scale-Invariant Signal-to-Distortion Ratio (SI-SDR) [89] gibi metriklerin referans gerektirmeyen elde edilmesini gerçekleştiren model [90] çıktıları da havuza dahil edilmiştir. Tablo 3.3'te tüm metadata detayları gösterilmiştir.

Tablo 3.3: Veri setine ait sık kullanılan metadatalar.

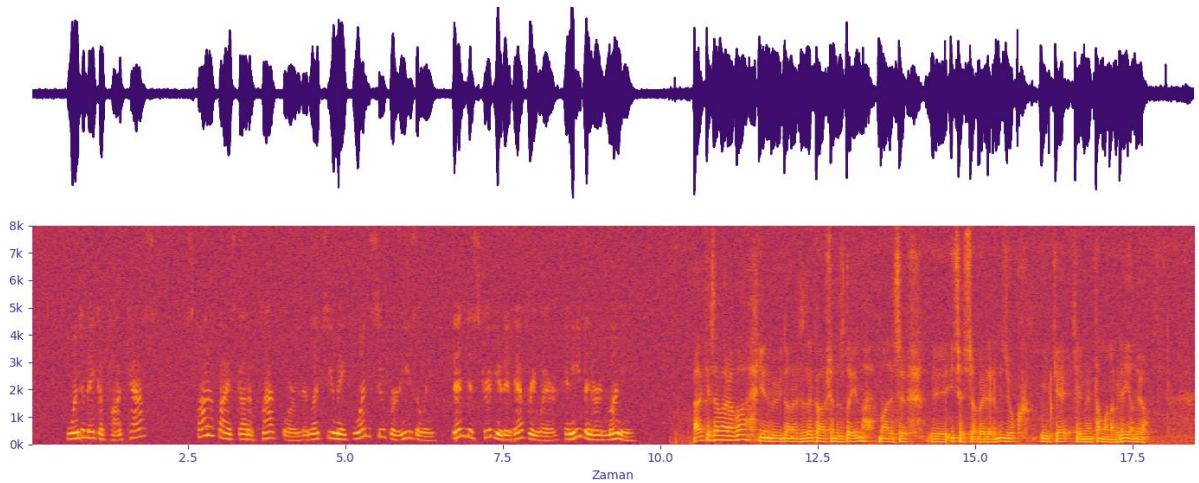
Metadata	**Açıklama**
Frekans	Orijinal ses dosyası frekansı.
Uzunluk	Ses parçasına ait uzunluk. (s)
Format	Dosya formatına dair detay.
Boyut	Dosyanın boyutu. (MB)
Hash	İlk 60 saniye için özet bir değer.
Konuşmacı	Konuşmacı kimlik numarası.
Cinsiyet	Konuşmacı cinsiyeti.
Ortalama Frekans	Frekans uzayındaki ortalamadır. (kHz)
Standart Sapma	Frekans uzayındaki standart sapma.
Medyan	Frekans uzayındaki medyan. (kHz)
İlk Çeyreklik	İlk çeyrekteki baskın frekansların medyanı. (kHz)
Üçüncü Çeyreklik	İlk üç çeyrek içerisindeki baskın frekansların medyanı. (kHz)
Çeyreklikler Arası	Üçüncü çeyrek ve ilk çeyrek arasındaki fark. (kHz)
Çarpıklık	Spektrumdaki asimetrinin bir ölçüsü.
Basıklık	Spektrumdaki dorukluluğun bir ölçüsü.
Spektral Entropi	Spektrumdaki enerji dağılımı.
Spektral Düzlük	Spektrumdaki düzlük.
Mod Frekansı	Frekansın modu veya baskın frekans.
Frekans Merkezi	Frekans ağırlık merkezi.
Ort. Temel Frekans	Temel frekans ortalaması.
Min. Temel Frekans	Temel frekansların minimumu.
Maks. Temel Frekans	Temel frekansların maksimumu.
Ort. Baskın Frekans	Baskın frekans ortalaması.
Min. Baskın Frekans	Baskın frekansların minimumu.
Maks. Baskın Frekans	Baskın frekansların maksimumu.
Baskın Frekans Aralığı	Baskın frekanslar arasındaki aralık.
Modülasyon İndisi	Modülasyon indisi.
PESQ	Konuşma kalitesinin geniş bant algısal tahmini.
STOI	Kısa-süreli amaç anlaşılırlığı.
SI-SDR	Ölçekle değişmeyen sinyal-bozulma oranı.

3.1.8. Ön İşleme Adımları Çevirilerin Toplanması ve Çapraz Doğrulama

Veri setimizin büyüklüğü göz önüne alındığında sıfırdan çevirilerin yapılması çok zorlu bir süreç olarak kendini göstermektedir. Bu nedenle el ile sıfırdan çeviriler yerine başarılı diğer modellerin kullanılmasıyla ilk doğrulamalar gerçekleştirilmiştir. Google Cloud Speech-toText [91], Microsoft Cognitive Services [92], Vosk [93] ve Whisper [15] gibi modellerin kullanımı ile özellikle kelime tabanlı kesinlik skorları alınmıştır. Kelime tabancı skorlar, farklı modellerin tek bir kaydı çapraz doğrulaması için kritik bir destek oluşturmuştur. Çevirilerdeki yabancı kelimelerin, kısaltmaların ve diğer olası beklenmedik telaffuzların daha sonrasında el ile çevirisinin gerekliliği ön görülmüştür. Bu noktada kitle kaynaklı olarak kullanılması ve çevirilerin bu şekilde doğrulanması noktasında farklı bir çalışma sürdürülmektedir.

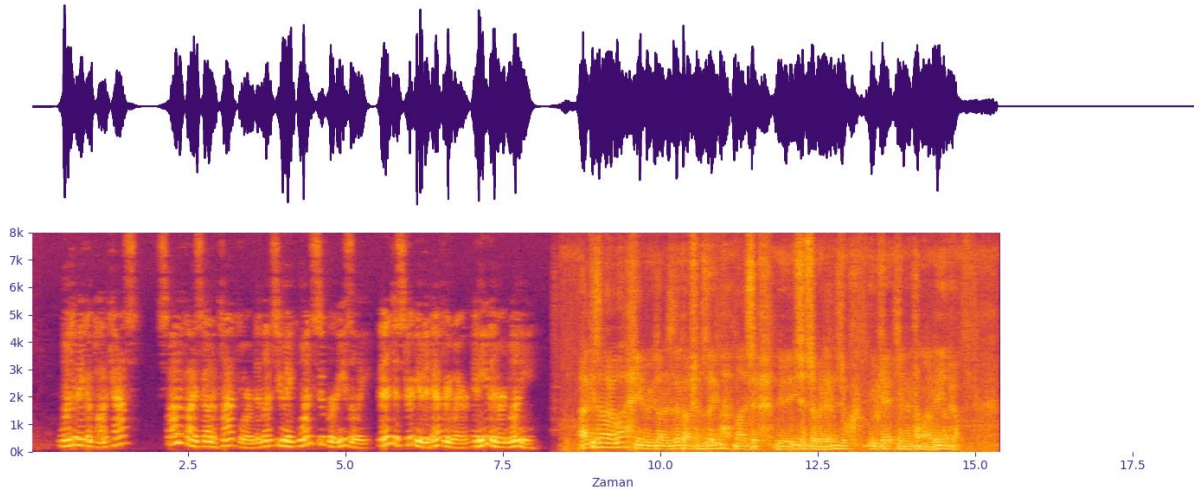
3.1.9. Veri Artırımı

Araştırmacılara daha farklı alt setler sunmak üzere nihai veri setimizde önceden gerçekleştirilmiş bazı veri artırımı derlemeleri de hazırlanmıştır. Bunun için klasik yöntemlerden olan gürültü ekleme, hız değişimi ve yankılama veya ses seviyesindeki modifikasyonları sağlayarak nihai veri setinin daha zorlu ve çeşitli hale getirilmesi adına veri artırımı planlanmıştır. Örnek sonuçlar gürültü ekleme, hız değişimi ve yankılama, ses seviye değişimi sırasıyla Şekil 8, 9 ve 10'da gösterilmiştir. Yanı sıra, 8 kHz olarak da literatür açığını desteklemek üzere özellikle sohbet verilerini akustik simülasyona [94] dahil edip gerçek bir telefon görüşmesine çok yakın alt kümenin oluşturulması da planlanmıştır.

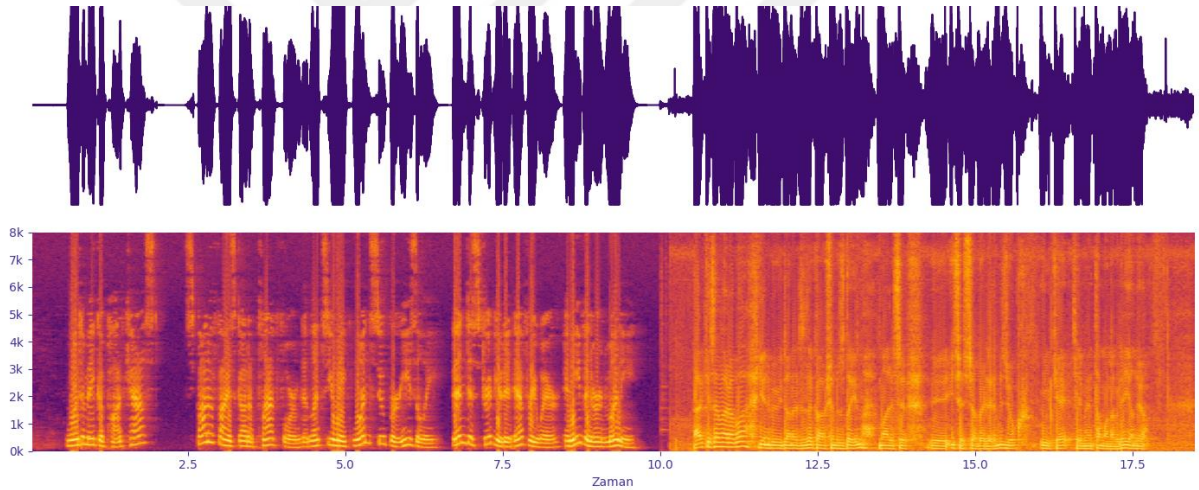


Şekil 3.6: Veri artırımı sonrası gürültü eklenmiş örnek. Üstte dalga formu altta ise

spektrogram.



Şekil 3.7: Veri artırımı sonrası hızı değiştirilmiş örnek. Üstte dalga formu altta ise spektrogram.



Şekil 3.8: Veri artırımı sonrası ses seviyesi değiştirilmiş örnek. Üstte dalga formu altta ise spektrogram.

3.2. Veri Setleri ve Modellerin Eğitimi

TURKSPEECH veri setinin otomatik etiketlenmiş bir kısmı ile Common Voice Türkçe veri seti kullanılarak, geleneksel ve yenilikçi modellerin yanında gözetimsiz yöntemler üzerinde de eğitimler yapılmıştır. Modellerin test edilmesi için Common Voice 15 versiyonlarına ait delta segmentler kullanılmıştır. Bunun yanında ise Common Voice 14 delta gözetimsiz yöntemlerin ince ayarı için tercih edilmiştir [28]. Veri setlerine ait detaylar Tablo 3.4'te verilmiştir.

Tablo 3.4: Kullanılan veri setlerine ait detaylar. (*: otomatik çeviri)

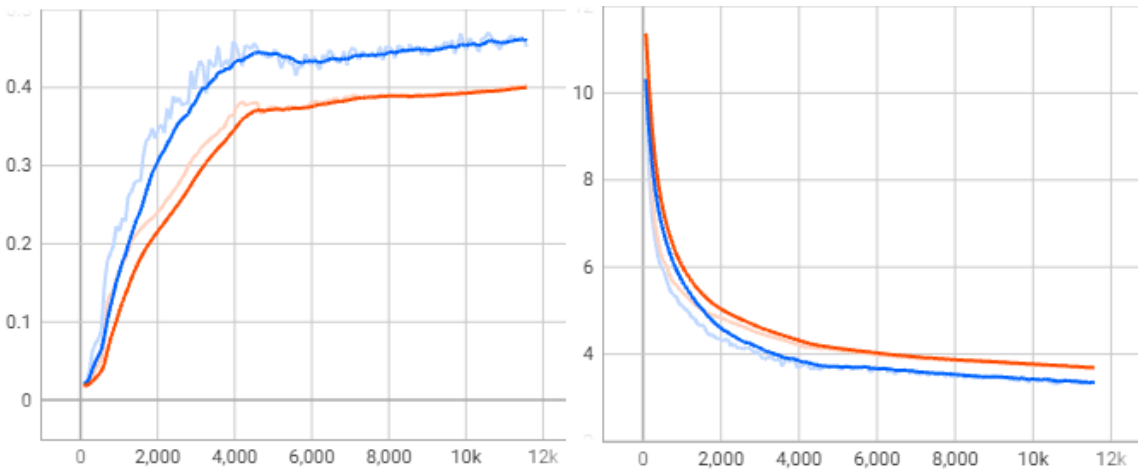
#	Kaynak	Eğitim (saat)	Test (saat)	Doğrulama (saat)
1	**TURKSPEECH**	200	10*	2
2	**Common Voice 15** - Ön Eğitim - İnce ayar	100 85 15	10 - -	1 - -

3.2.1. Kaldi ile Yapılan Eğitimler

Geleneksel yöntemleri temsilen Kaldi-ASR kullanılmıştır [66]. Kaldi içerisindeki özgün sinir ağı mimarisini taban alan Librispeech baz alınmıştır. Kaldi gözetimli öğrenme gerçekleştirdiğinden bu eğitimde 2 numaralı veri setindeki 100 saatlik parça eğitim ve 10 saatlik parça test için kullanılmıştır.

3.2.2. Wav2vec2 ile Gözetimsiz Öğrenme

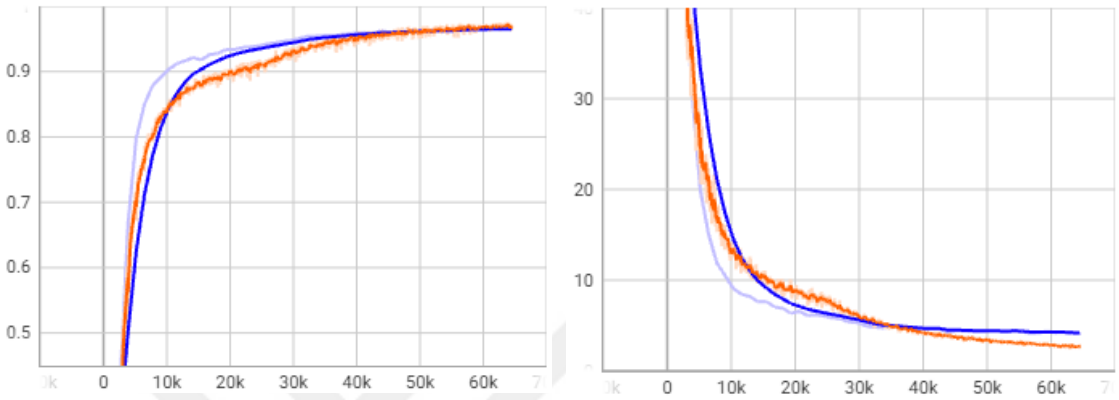
Gözetimsiz bir öğrenme gerçekleştirmek için wav2vec2 tercih edilmiştir [77]. Eğitim kümesi olarak 1 ve 2 numaralı veri setleri kullanılmıştır. İlk olarak 1 numaralı veri seti tümüyle kullanılarak ön eğitim ve arkasından ise 2 numaralı veri setinin ilgili kırılımı ile ince ayar uygulanmıştır. Bunun yanı sıra farkı gözlemlemek üzere 2 numaralı veri setine ait ön eğitim ve aynı ince ayar kırılımı kullanılarak tekrar eğitim yapılmıştır. İlk eğitime ait grafikler Şekil 3.9'da gösterilmiştir. Mavi renk doğrulama, turuncu renk ise eğitim kümesini temsil etmektedir.



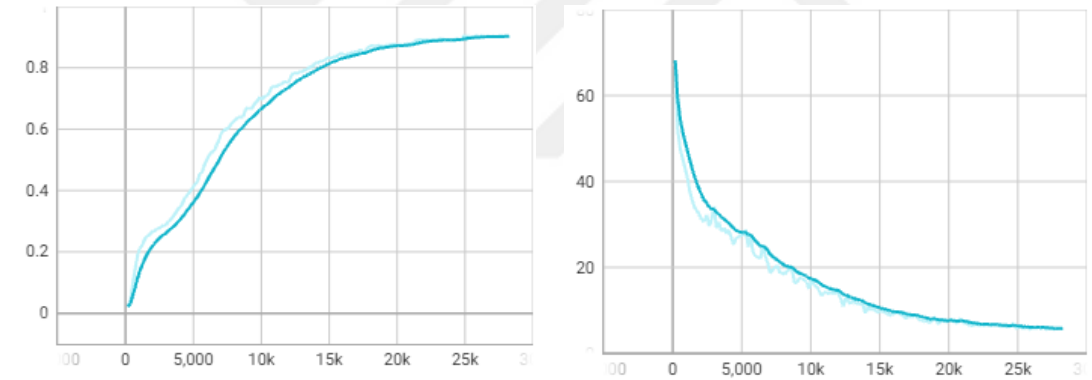
Şekil 3.9: Wav2vec2 eğitime ait bir görsel. Solda eğitim başarısı, sağda hatası.

3.2.3. Uçtan-uca Konuşma Tanıma: Espnet - Sherpa

Uçtan-uca konuşma tanıma deneylerinde doğrulanmış veri seti olan Common Voice tercih edilmiştir. Espnet ve Sherpa için iki farklı Transformer bu eğitimlerde tercih edilmiştir. Espnet için Şekil 3.10 ve 3.11'de sırasıyla Conformer ve Transformer eğitim grafikleri verilmiştir [24, 96, 97].



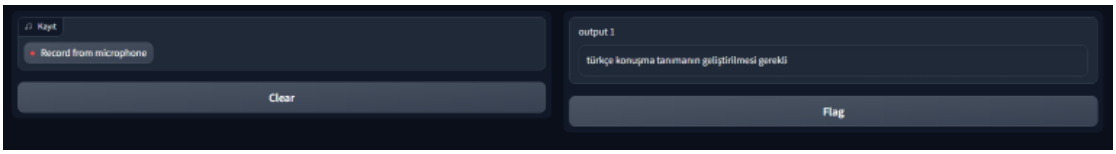
Şekil 3.10: Espnet Conformer eğitimi. Solda başarı sağda hata değişimi.



Şekil 3.11: Espnet Transformer eğitimi. Solda başarı sağda hata değişimi.

3.3. Gerçek Zamanlı Konuşma Tanıma Uygulaması

Gerçek zamanlı konuşma tanıma uygulaması için Python kullanılarak gradio isimli kütüphane yazılımı ile basit bir konuşma tanıma uygulaması geliştirilmiştir. Gerçek zamanlı ses akışı bu aracı program ile sağlanarak ses parçalarının modele verilmesi sağlanmıştır. Silero VAD ile uygulamanın gürültünün etkisi azaltılmıştır [83].



Şekil 3.12: Basit bir gradio uygulaması.

4. BULGULAR VE TARTIŞMA

Bu çalışma sonucunda [40, 41] çalışmalarına benzer şekilde veri seti derlemesi elde edilmiştir. Kolobov ve ark. 10 saatlik Türkçe verinin derlenmesini gerçekleştirmişledir [40], TURKSPEECH ise bu çalışmaya kıyasla yalnızca haber kaynaklarını kullanmayıp gerçek dünya verisine daha yakın sohbet konuşmalarına da yer vermiştir. Çeşitlik ve uzunluk açısından veri setimiz toplamda 6290 saat uzunluğuyla bildiğimiz kadarıyla derlenmiş Türkçe veri setleri arasında en büyüğü konumundadır. Ayrıca veri setimizin metadata zenginliği de benzeri çalışmalara kıyasla daha fazla olmuştur [40, 41]. Zang, Lv ve ark. [41] çalışmamıza yakın metotlar ile 10000 saati doğrulanmış olarak 20000 saatten fazla ses kaydını literatüre kazandırmıştır. Mandarin dilindeki bu veri seti üzerinde yapılan çalışmalar bizim de çalışmamızda ön gördüğümüz gibi ilgili dil için konuşma tanımanın başarısı artırmış görünmektedir. [41] ile rapor edilen 8,8 kelime hata oranı skoru Lee ve ark. tarafından 7,19 olarak güncellenmiştir. Bir yıl sonrasında ise Gao ve ark. 60 bin saatlik doğrulanmış Mandarin konuşma verisi ile ilgili veri setindeki kelime hata oranını 6,9'a düşürmüştür [95]. Gao ve ark. bu çalışmasında Wenet veri setini [41] yalnızca test amaçlı kullanmışlardır. Sonuçlar gösteriyor ki yüksek miktarda ve çeşitlilikte ses verisi uçtan-uca konuşma tanıma modellerinin gelişiminde büyük bir fark yaratmaktadır.

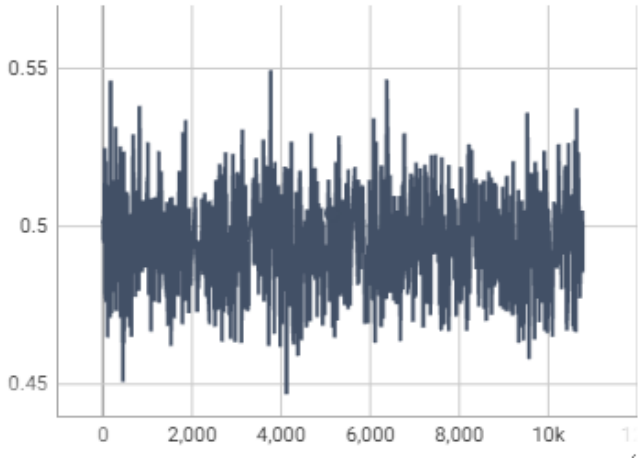
4.1. TURKSPEECH Veri Seti için İstatistik ve Bulgular

Bu çalışmada derlemesi gerçekleştirilen veri setine ait istatistik bilgiler Tablo 4.1'de paylaşılmıştır. Veri setinin büyük boyutlu oluşu beraberinde zorluklar getirse de daha detaylı olarak incelenmesi ve çevirilerinin tekrar doğrulanması üzerinde gelecek adımlar planlanmıştır. Bu noktada gözetimsiz öğrenim metotları ile el ile etiketleme yöntemini birleştirecek farklı bir yaklaşım da umut verici sonuçlar ortaya koyabilir. İlgili veri setinin veri artırımı aşamasının tamamlanmasıyla Türkçe özelinde geliştirmekte olduğumuz konuşma tanıma modeline dair sonuçlar referans çalışma olarak paylaşılması düşünülmektedir. Bu çalışmanın hem veri hem de veri seti hazırlığında sunduğu yaklaşımlar ve problemlere dair çözümler ile, ilgili araştırmacılara hız kazandırıcı bir kaynak olarak katkı sağlaması amaçlanmıştır.

Tablo 4.1: TURKSPEECH Veri seti istatistikleri.

	Süre (saat)	Parça (adet)	Boyut (GB)
Orijinal	63244	107503	4449
Wav	63244	107503	38418
Wav 16k	63244	107503	6874
Wav 16k VAD	6290	1760899	100

Öte yandan TURKSPEECH veri seti boyutu itibarı her ne kadar gözetimsiz öğrenme metotları için uygun olsa da Wav2vec2 gibi büyük mimarilerle gerçekleştirdiğimiz eğitimlerin tamamlanma ve doyuma ulaşma süresi oldukça zorlayıcı olmuştur. Veri setinin çok farklı akustik ortamları bir araya getirmesi ve ses çeşitliliği açısından yüksek miktarlar sunmasının bu problemin oluşmasında etkili olduğunu düşünülmektedir. Diğer taraftan ise Wav2vec2 gibi gözetimsiz metotlar için kullanılan donanım gücüne bu çalışmada erişilememesi de eğitim grafiklerinin diğer çalışmalardaki gibi iyi ve belirgin başarıya ulaşamamasına neden olduğunu düşünülmektedir [76]. Şekil 4.1'de benzeri bir durum için görsel verilmiştir (bu eğitim yaklaşık bir hafta sürmüştür). Bu nedenlerle deneylerimizde bu veri setinin bir alt kırılımı olarak seçilen 200 saatlik en temiz küme tercih edilmiştir.



Şekil 4.1: TURKSPEECH tamamının Wav2vec2 ile ön eğitimine ait hata grafiği.

5. SONUÇLAR

Eğitimi gerçekleştirilen modellere ait sonuçların tümü ilgili modellere ait eğitimler süresince görülmemiş örneklerden oluşan Common Voice içerisindeki 10 saatlik test kümesi ve TURKSPEECH veri seti içerisinden seçilen otomatik çevirilere ait 10 saatlik test seti kırılımı üzerinden yapılmıştır. En iyi sonuçlar Tablo 5.1'de paylaşılmıştır.

Tablo 5.1: Modellere ait kelime hata oranları.

Model Mimarisi	**Common Voice Test (wer)**	**TURKSPEECH Test (wer)**
Kaldi Librispeech	11,2	30,1
Espnet Conformer	**10,8**	**28,2**
Espnet Transformer	13,4	34,9
Sherpa Transducer	50,0	63,4
Wav2Vec2 Base	18,3	29,0

Yapılan testler sonucunda Tablo 5.1'de görüldüğü gibi TURKSPEECH veri seti üzerinden alınan kelime hata oranları Common Voice test kümesine göre nispeten yüksek kalmakta ve %28 altına düşmemektedir. Bu duruma neden olan temelin otomatik çeviriler olduğu düşünülmektedir. Zira sonuçlar arasında aslında hipotezlerin doğru fakat otomatik oluşturulmuş referans çevirilerin yanlış olduğu örnekler gözlemlenmiştir. Bu nedenlerle TURKSPEECH veri seti için daha detaylı bir çeviri yaklaşımının önemi tekrar kendini göstermiş ve gelecek çalışmalarda ele alınmak üzere planlanmıştır. Bu veri seti için en iyi sonuçlar ise benzeri veri grubu ile eğitilen Wav2Vec2 ve Espnet Conformer üzerinden hesaplanmıştır.

Öte yandan Kaldi ile gerçekleştirilen yeni nesil model ise doğrulanmış veri kümesi olan Common Voice üzerinde Espnet Conformer sonrasında en iyi sonuçları üretebilmiştir. Bu test kümesinde ise Wav2Vec2 ile alınan %18,3 oranındaki kelime hata oranı gözetimsiz

bir için umut vadeden bir sonuç olmuştur. İlgili model için 200 saat yerine 6000 saatlik kısmının kullanılması ile daha iyi sonuçların da üretilebileceği ön görülmektedir. Diğer yandan Espnet Transformer modeli ise esasında "streaming" yapısı ile eğitildiği ve gerçek zamanlı hipotez üretmeye odaklı olması sebebiyle Conformer altyapısına göre daha yüksek hata oranı üretmiştir. Testlerimizde düşük performans gösteren model ise Sherpa ile denediğimiz özgün Transducer altyapısına sahip model oluşmuştur. Bu model için Türkçe dil desteğinin kısıtlı olması sebebiyle sonuçların diğer modellere kıyasla düşük geldiği düşünülmektedir. En başarılı model olarak ise Espnet Conformer modeli kendini göstermektedir. Bu modele ait bazı örnek hipotezler Tablo 5.2'de verilmiştir.

Tablo 5.2: Espnet Conformer modeli için bazı hipotezler.

Referans Çeviri	Hipotez Çeviri
ŞİMDİ BİR KAHVE ÇOK İYİ GİDERİ	ŞIMDI BIR KAHVE ÇOK IYI GIDERDI
BUNUN NASIL YAPILDIĞINI BİLMİYORUM	BUNUN NASIL YAPILDIĞINI BILMIYORUM
PEYNİRLİ VE SUCUKLU TOST SİPARİŞ ETTİM	PEYNIRLI VE ÇUCUKLU DOST PARIS ETTIM

KAYNAKLAR

[1] S. Ramakrishnan, Modern Speech Recognition: Approaches with Case Studies. InTech, 2012.

[2] L. Rabiner and B. H. Juang, Fundamentals of speech recognition. Prentice-Hall, 1993.

[3] Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A. R., Jaitly, N., Kingsbury, B., "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups." IEEE Signal Processing Magazine, 29(6), pp. 82-97, 2012.

[4] S. Hochreiter and S. Jürgen, "Long short-term memory." Neural computation, 9(8), pp. 1735-1780. 1997.

[5] Graves, A., Fernández, S., Gomez, F., & Schmidhuber, J., "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks." In Proc. of the 23rd international conference on Machine learning. 2006.

[6] Sak, H., Senior, A., Rao, K., & Beaufays, F., "Fast and accurate recurrent neural network acoustic models for speech recognition." In Proc. Interspeech, pp. 1468-1472, 2015.

[7] A. Graves and J. Navdeep, "Towards end-to-end speech recognition with recurrent neural networks." In Proc. International conference on machine learning, 2014.

[8] https://pnwaudiology.com/hearing-loss/overview/how-we-hear/

[9] Vasif V. Nabiyev ve Ergün Yücesoy, VQ Yöntemiyle Konuşmacı CinsiyetininBelirlenmesi, Turkish Journal of Computer and Mathematics Education, vol 1,1 (2009) 35-47

[10] Noelia Alcaraz Meseguer, Speech Analysis for Automatic Speech Recognition,Norwegian University of Science and Technology Department of Electronics and Telecommunications, (2009)

[11] Steven W. Smith, The Scientist and Engineer's Guide to Digital Signal Processing, (1997)

[12] Davis S., Mermelstein P., Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, IEEE Transactions on Acoustics, Speech, and Signal Processing, 28(4), 357-366, August 1980.

[13] Ramesh A., Pavlov M., Goh G., Gray S., Voss C., Radford A., Chen M., Sutskever I., Zero-shot text-to-image generation, Proceedings of the 38th International Conference

on Machine Learning, PMLR, 139, 8821-8831, 2021.

[14] Tom B., Benjamin M., Nick R., Melanie S., Jared D., Prafulla D., Arvind N., Pranav S., Girish S., Amanda A., Sandhini A., Ariel H., Gretchen K., Tom H., Rewon C., Aditya R., Daniel Z., Jeffrey W., Clemens W., Chris H., Mark C., Eric S., Mateusz L., Scott G., Benjamin C., Jack C., Christopher B., Sam M., Alec R., Ilya S., Dario A., Language models are few-shot learners, Language models are few-shot learners, Advances in Neural Information Processing Systems, 33, 1877-1901, 2023.

[15] Radford A., Kim J. W., Xu T., Brockman G., McLeavey C., Sutskever I., Robust Speech recognition via large-scale weak supervision, International Conference on Machine Learning, ICML, 28492-28518, 2023.

[16] Paul, D. B., & Baker, J. (1992). The design for the Wall Street Journal-based CSR corpus. In Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992.

[17] Panayotov, V., Chen, G., Povey, D., & Khudanpur, S., "Librispeech: an ASR Corpus Based on Public Domain Audio Books." In Proc. ICASSP, 2015.

[18] Bu, H., Du, J., Na, X., Wu, B., & Zheng, H. (2017, November). Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline. In 2017 20th conference of the oriental chapter of the international coordinating committee on speech databases and speech I/O systems and assessment (O-COCOSDA) (pp. 1-5). IEEE.

[19] Du, J., Na, X., Liu, X., & Bu, H. (2018). Aishell-2: Transforming mandarin asr research into industrial scale. arXiv preprint arXiv:1808.10583.

[20] Rousseau, A., Deléglise, P., & Esteve, Y. (2012, May). TED-LIUM: an Automatic Speech Recognition dedicated corpus. In LREC (pp. 125-129).

[21] Lüscher, C., Beck, E., Irie, K., Kitza, M., Michel, W., Zeyer, A.,"RWTH ASR Systems for LibriSpeech: Hybrid vs Attention--w/o Data Augmentation." In Proc. Interspeech, pp. 231-235, 2019.

[22] Karita, S., Chen, N., Hayashi, T., Hori, T., Inaguma, H., Jiang, Z., "A comparative study on transformer vs rnn in speech applications." In Proc. ASRU, 2019.

[23] Gulati, A., Qin, J., Chiu, C. C., Parmar, N., Zhang, Y., Yu, J., "Conformer: Convolution-augmented transformer for speech recognition." In Proc. Interspeech, pp. 5036-5040, 2020.

[24] Guo, P., Boyer, F., Chang, X., Hayashi, T., Higuchi, Y., Inaguma, H., ... & Zhang,

Y. (2021, June). Recent developments on espnet toolkit boosted by conformer. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 5874-5878). IEEE.

[25] wer_are_we, Gabriel Synnaeve, Github. https://github.com/syhw/wer_are_we. Erişim tarihi Temmuz 1, 2023.

[26] Papers with Code, https://paperswithcode.com/, Erişim tarihi Temmuz 1, 2023.

[27] OpenSLR, Open Speech and Language Resources, https://openslr.org. Erişim tarihi Temmuz 1, 2023.

[28] Mozilla Common Voice, Mozilla Foundation, https://commonvoice.mozilla.org/. Erişim tarihi Temmuz 1, 2023.

[29] Mozilla Foundation Blog, https://foundation.mozilla.org/en/blog/mozillas-common-voice-dataset-reaches-100-languages/. Yayın tarihi Eylül 15, 2022. Erişim tarihi Temmuz 1, 2023.

[30] The Linguistic Data Consortium, University of Pennsylvania, https://www.ldc.upenn.edu/, Erişim tarihi Temmuz 1, 2023

[31] Salor Ö., Ciloglu T., Pellom B., Demirekler M., Middle East Technical University Turkish Microphone Speech v 1.0 LDC2006S33. Philadelphia: Linguistic Data Consortium, 2006.

[32] Saraclar. M., Turkish broadcast news speech and transcripts LDC2012S06. Philadelphia, Linguistic Data Consortium, Web Download, 2012.

[33] Martin A. F., Greenberg C. S., The 2009 NIST language recognition evaluation, Odyssey 2010: The Speaker and Language Recognition Workshop, Czech Republic, June 28 - July 1, 2010.

[34] Andresen J., Bills A., Dubinski E., Fiscus J., Gillies B., Harper M. T., Rytting A., IARPA Babel Turkish Language Pack, IARPA-babel105bv0. 5 LDC2016S10. Web Download. Philadelphia: Linguistic Data Consortium, 2016.

[35] Jones K., Multi-Language Conversational Telephone Speech 2011 Turkish LDC2017S09. Web Download. Philadelphia: Linguistic Data Consortium, 2017.

[36] Greenberg C.S., Martin A.F., Przybocki M., The 2011 NIST language recognition evaluation, 13th Annual Conference of the International Speech Communication Association 2012, INTERSPEECH 2012. 1. 34-37. 2012.

[37] Büyük O., A New Database For Turkish Speech Recognition On Mobile Devices

And Initial Speech Recognition Results Using The Database, Pamukkale University Journal of Engineering Sciences-Pamukkale Universitesi Muhendislik Bilimleri Dergisi, 24(2), 180-184, 2018.

[38] Fakhan E., Automatic Speech Recognition System Adaptation For Spoken Lecture Processing, PhD. Thesis, Boğaziçi University, İstanbul, Turkey, 2021.

[39] Oyucu S., Polat H., Sever H., Web Service-Based Turkish Automatic Speech Recognition Platform, In Proc. International Congress on Human-Computer Interaction, Optimization and Robotic Applications, 1-5, 2020.

[40] Kolobov R., Okhapkina O., Omelchishina O., Platunov A., Bedyakin R., Moshkin V., Menshikov D., Mikhaylovskiy N., Mediaspeech: multilanguage asr benchmark and dataset. arXiv preprint arXiv:2103.16193, 2021.

[41] Zhang B., Lv H., Guo P., Shao Q., Yang C., Xie L., Xu X., Bu H., Chen X., Zeng C., Wu D., Peng Z., Wenetspeech: A 10000+ hours multi-domain mandarin corpus for speech recognition, International Conference on Acoustics, Speech and Signal Processing, ICASSP, 6182-6186, 2022 2022.

[42] Carki, K., Geutner, P., & Schultz, T., "Turkish LVCSR: towards better speech recognition for agglutinative languages." In Proc. ICASSP, 2000.

[43] Sak, H., Saraçlar, M., & Gungor, T., "Morpholexical and discriminative language models for Turkish automatic speech recognition." IEEE transactions on audio, speech, and language processing, 20(8), pp. 2341-2351, 2012.

[44] Sak, H., Saraclar, M., & Güngör, T., "Morphology-based and sub-word language modeling for Turkish speech recognition." In Proc. ICASSP, 2010.

[45] Arısoy, E., Dutağacı, H., & Arslan, L. M., "A unified language model for large vocabulary continuous speech recognition of Turkish." Signal Processing, 86(10), pp. 2844-2862, 2006.

[46] O. Büyük, "Sub-world language modelling for Turkish." PhD. Thesis, Sabancı University, İstanbul, Turkey, 2005.

[47] Bayer, A. O., Ciloglu, T., & Yondem, M. T., "Investigation of different language models for Turkish speech recognition." In Proc. IEEE 14th Signal Processing and Communications Applications, 2006.

[48] E. Arısoy, "Statistical and discriminative language modeling for Turkish large vocabulary continuous speech recognition." PhD. Thesis, Boğaziçi University, İstanbul,

Turkey, 2009.

[49] R. S. Arslan, and N. Barışçı. "A detailed survey of Turkish automatic speech recognition." Turkish Journal of Electrical Engineering & Computer Sciences, 28(6), pp. 3253-3269. 2020.

[50] B. Tombaloğlu, and H. Erdem. "Turkish Speech Recognition Techniques and Applications of Recurrent Units (LSTM and GRU)." Gazi University Journal of Science, 34(4), pp. 1035-1049. 2021.

[51] U. A Kimanuka, and O. Büyük. "Turkish speech recognition based on deep neural networks." Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi, 22, pp. 319-329, 2018.

[52] B. Tombaloğlu, and H. Erdem. "Deep Learning Based Automatic Speech Recognition for Turkish." Sakarya University Journal of Science, 24(4), pp. 725-739, 2020.

[53] E. Fakhan and E. Arısoy, "Domain Adaptation Approaches for Acoustic Modeling." In Proc. IEEE 28th Signal Processing and Communications Applications Conference, 2020.

[54] B. Asefisaray, "Uçtan-uca konuşma tanıma modeli: Türkçedeki deneyler." PhD. Thesis, Hacettepe University, Ankara, Turkey, 2018.

[55] Salor Ö., Pellom B. L., Ciloglu T., Demirekler M., Turkish speech corpora and recognition tools developed by porting SONIC: Towards multilingual speech recognition, Computer Speech & Language, 21(4), 580-593, 2007.

[56] Polat H., Oyucu S., Building a Speech and Text Corpus of Turkish: Large Corpus Collection with Initial Speech Recognition Results, SYMMETRY-BASEL, 12(2), 290, 2020.

[57] Arisoy E., Can D., Parlak S., Sak H., Saraçlar M., Turkish broadcast news transcription and retrieval, IEEE Transactions on Audio, Speech, and Language Processing, 17(5), 874-883, 2009.

[58] Arısoy E., Saraçlar M., Turkish Broadcast News Transcription Revisited, The IEEE 28. Sinyal İşleme ve İletişim Uygulamaları Konferansı, SIU, 1-4, 2018.

[59] Arısoy E., Sak H., Saraçlar M., Language modeling for automatic Turkish broadcast news transcription, INTERSPEECH 2007: 8th Annual Conference of the International Speech Communication Association, 2748-2751, 2007.

[60] Gokay R., Yalcin H., Improving low resource Turkish speech recognition with data augmentation and TTS,16th International Multi-Conference on Systems, Signals & Devices, SSD, 357-360, March, 2019.

[61] Ren Z., Yolwas N., Wang H., Slamu W., Exploring Turkish Speech Recognition via Hybrid CTC/Attention Architecture and Multi-feature Fusion Network. arXiv preprint arXiv:2303.12300, 2023.

[62] Mercan O. B., Cepni S., Tasar D. E., Ozan S., Performance Comparison of Pre-trained Models for Speech-to-Text in Turkish: Whisper-Small and Wav2Vec2-XLS-R-300M. arXiv preprint arXiv:2307.04765, 2023.

[63] Speech Recognition on Common Voice Turkish, Papers With Code, https://paperswithcode.com/sota/speech-recognition-on-common-voice-turkish. Erişim tarihi Temmuz 1, 2023.

[64] K. F. Lee, H. W. Hon, and R. Reddy, "An overview of the SPHINX speech recognition system." IEEE Transactions on Acoustics, Speech, and Signal Processing, 38(1), pp. 35-45, 1990.

[65] S. J. Young, "The HTK hidden Markov model toolkit: Design and philosophy." Entropic Cambridge Research Laboratory, 2, pp. 2-44, 1993.

[66] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Vesely, K., "The Kaldi speech recognition toolkit." In Proc. IEEE 2011 workshop on automatic speech recognition and understanding, 2011.

[67] Palaz, H., Kanak, A., Bicil, Y., Dogan, M. U., & Islam, T., "TREN-Turkish speech recognition platform." In Proc. 13th European Signal Processing Conference, 2005.

[68] H. Dutağacı, "Statistical language models for large vocabulary Turkish speech recognition." MS Thesis, Department of Computer Engineering, Boğaziçi University, İstanbul, 2002.

[69] E. Dikici, "Supervised, Semi-Supervised And Unsupervised Methods In Discriminative Language Modeling For Automatic Speech Recognition." PhD Thesis, Boğaziçi University, İstanbul, Turkey, 2016.

[70] M. J. Ahmed, "Çağrı merkezleri için derinöğrenme tabanlı interaktif konuşma tanıma." MS Thesis, Selçuk University, Fen Bilimleri Enstitüsü, Konya, Turkey, 2020.

[71] R. S. Arslan, "Development of output correction methodology for turkish speech recognition and design of a recurrent neural network." PhD. Thesis, Gazi University,

Ankara, Turkey, 2020.

[72] H. Kutucu, "Derin öğrenme algoritmaları kullanarak bir konuşma tanıma uygulaması." MS Thesis, Sakarya Uygulamalı Bilimler Üniversitesi, Sakarya, Turkey, 2020.

[73] S. Oyucu, "Türkçe Konuşma Tanıma Sistemleri için Derin Öğrenme Tabanli Modellerin Geliştirilmesi" PhD. Thesis, Hacettepe University, Ankara, Turkey, 2020.

[74] B. Tombaloğlu, "Automatic speech recognition and sign language translation for Turkish." PhD. Thesis, Sakarya University, Sakarya, Turkey, 2021.

[75] Yu, Chongchong, et al. "Acoustic modeling based on deep learning for low-resource speech recognition: An overview." IEEE Access 8 (2020): 163829-163843.

[76] Babu, Arun, et al. "XLS-R: Self-supervised cross-lingual speech representation learning at scale." arXiv preprint arXiv:2111.09296 (2021).

[77] Baevski, Alexei, et al. "wav2vec 2.0: A framework for self-supervised learning of speech representations." Advances in neural information processing systems 33 (2020): 12449-12460.

[78] Kuchaiev, Oleksii, et al. "Nemo: a toolkit for building ai applications using neural modules." arXiv preprint arXiv:1909.09577 (2019).

[79] Xu, Qiantong, et al. "Self-training and pre-training are complementary for speech recognition." ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021.

[80] Hsu, Wei-Ning, et al. "Robust wav2vec 2.0: Analyzing domain shift in self-supervised pre-training." arXiv preprint arXiv:2104.01027 (2021).

[81] Conneau, Alexis, et al. "Unsupervised cross-lingual representation learning for speech recognition." arXiv preprint arXiv:2006.13979 (2020).

[82] Google WebRTC. Google LLC, https://webrtc.org/. Temmuz 1, 2023.

[83] Silero Team, Silero VAD: pre-trained enterprise-grade Voice Activity Detector (VAD), Number Detector and Language Classifier, https://github.com/snakers4/silero-vad, 2021.

[84] Bredin H., Yin R., Coria J., Gelly G., Korshunov P., Lavechin M., Fustes D., Titeux H., Bouaziz W., Gill M. P., Pyannote audio: neural building blocks for speaker diarization, IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, 124-7128, 2020.

[85] Bredin H., Laurent A., End-to-end speaker segmentation for overlap-aware resegmentation, arXiv preprint arXiv:2104.04045, 2021.

[86] Kalaycı E. E., Doğan B., Gender Recognition by Using Acoustic Features of Sound with Deep Learning and Data Mining Methods, Innovations in Intelligent Systems and Applications Conference, ASYU, 1-4, 2020.

[87] Union I. T., Wideband extension to recommendation p. 862 for the assessment of wideband telephone networks and speech codecs. International Telecommunication Union, Recommendation P, 862, 2007.

[88] Taal C. H., Hendriks R. C., Heusdens R., Jensen J., A short-time objective intelligibility measure for time-frequency weighted noisy speech, In 2010 IEEE international conference on acoustics, speech and signal processing, 4214-4217, 2010.

[89] Le Roux J., Wisdom S., Erdogan H., Hershey J., SDR–half-baked or well done?, IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, 2019.

[90] Kumar A., Tan K., Ni Z., Manocha P., Zhang X., Henderson E., Xu B., Torchaudio-Squim: Reference-Less Speech Quality and Intelligibility Measures in Torchaudio, IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, 1-5, 2023.

[91] Google Cloud Speech-to-Text, Google Inc. https://cloud.google.com/speech-to-text, Erişim tarihi Temmuz 1, 2023.

[92] Microsoft Cognitive Speech Services, Microsoft Corp., https://azure.microsoft.com/en-us/products/cognitive- services/speech-services/, Erişim tarihi Temmuz 1, 2023.

[93] Vosk Speech Recognizer, Alpha Cephei, https://alphacephei.com/vosk/, Erişim tarihi Temmuz 1, 2023.

[94] Ferras M., Madikeri S., Motlicek P., Dey S., Bourlard H., A large-scale open-source acoustic simulator for speaker recognition, IEEE Signal Processing Letters 23(4), 527-531, 2016.

[95] Gao Z., Li Z., Wang J., Luo H., Shi X., Chen M., Li Y., Zuo L., Du Z., Xiao Z., Zhang S., FunASR: A Fundamental End-to-End Speech Recognition Toolkit. arXiv preprint arXiv:2305.11013, 2023

[96] Sherpa, https://github.com/k2-fsa/sherpa, Erişim tarihi 1Temmuz 2023

[97] Icefall, https://github.com/k2-fsa/icefall, Erişim tarihi 1Temmuz 2023

ÖZGEÇMİŞ

Eğitim

02.2020 – Halen	**Yüksek Lisans, Bilgisayar Mühendisliği**, Marmara Üniversitesi, Türkiye Tez: Gerçek Zamanlı Türkçe Konuşma Tanıma
09.2014 – 06.2018	**Lisans, Bilgisayar Mühendisliği**, Karadeniz Teknik Üniversitesi, Türkiye Tez: Sesli Satranç, Ses Komutlarının Satranç Hamlesine Dönüştürülmesi

Deneyim

12.2021 – Halen	**Ar-Ge Mühendisi, Validsoft**
08.2019 – 12.2021	**Ar-Ge Mühendisi, Turkcell Global Bilgi**
07.2018 – 08.2019	**Serbest Oyun Geliştiricisi**
07.2018 – 09.2018	**Ar-Ge Stajı, Mavialp Bilgi Teknolojileri**
06.2017 – 08.2017	**Yüz Tanıma Staj, Bilgisayar Mühendisliği, KTÜ**
07.2016 – 08.2016	**Görüntü İşleme Temelleri Staj, Biyoistatistik Mühendisliği, KTÜ**