# Konuşma Tanıma ve Ses İşleme

## Yapay Zekâ ve Veri Dersleri Rehberi · Kısım III: Sinyal, Görüntü ve Ses

Sesin sayısallaşması | Spektrogram ve MFCC | Akustik model | Dil modeli | Konuşmacı ve olay | Komşu dersler

Sercan KÜLCÜ · Giresun Üniversitesi · Bilgisayar Mühendisliği

© Sercan Külcü. Tüm hakları saklıdır. · Sunumun etkileşimli HTML sürümü:
https://sercankulcu.github.io/files/ai_courses/sunumlar/Bolum_19_Konusma_ve_Ses_Isleme.html

GİRİŞ

# Ders Hangi Soruyu Sorar?

> **Ses dalgasından sözcükler, konuşmacı ve olaylar nasıl çıkarılır?**

Konuşma tanıma ve ses işleme, mikrofondan gelen ses dalgasını çözümleyip ondan sözcükleri, konuşmacının kimliğini ya da ortamdaki olayları çıkarmayı inceler. Ses önce zaman-frekans özniteliklerine çevrilir; **akustik model** ve dil modeli birlikte en olası metni bulur.

## Bir bakışta

| | | |
|---|---|---|
| **YAKLAŞIM**<br>akustik sinyal ve dil modeli | **GİRDİ**<br>ses dalgası | **ÇIKTI**<br>metin, konuşmacı, ses olayı |
| **BAŞARI ÖLÇÜTÜ**<br>sözcük hata oranı (WER) | **MATEMATİK YÜKÜ**<br>orta-yüksek | **VERİ İHTİYACI**<br>çok saatlik kayıt |

GİRİŞ

# Kısa Tarihçe

**1952**
Bell Laboratuvarları'nın Audrey sistemi tek bir konuşmacının söylediği rakamları tanır.

**1962**
IBM Shoebox, on altı İngilizce sözcüğü ve rakamları tanıyarak tanıtılır.

**1970'ler**
DARPA destekli projelerde Carnegie Mellon'ın Harpy sistemi yaklaşık bin sözcük tanır.

**1980'ler**
Saklı Markov modelleri (HMM) konuşma tanımanın standart yöntemi olur.

**2012**
Derin sinir ağları akustik modellemede Gauss karışımlarının yerini almaya başlar.

**2022**
OpenAI'ın Whisper modeli yaklaşık 680 bin saatlik çok dilli veriyle uçtan uca eğitilir.

DERS

# Temel Kavramlar

**Örnekleme**
Ses dalgasını saniyede binlerce kez ölçüp sayıya çevirmek.

**Spektrogram**
Kısa zamanlı Fourier dönüşümüyle zaman-frekans görüntüsü.

**MFCC**
İnsan kulağının frekans algısına uyarlanmış öznitelikler.

**Akustik model**
Ses parçasından ses birimini (fonem) tahmin eder.

**Dil modeli**
Hangi sözcük dizisinin daha olası olduğunu söyler.

**Kod çözme**
Işın aramasıyla en olası sözcük dizisini bulmak.

**Konuşmacı tanıma**
Sesin kime ait olduğunu doğrulamak ya da bulmak.

**Sözcük hata oranı**
WER: ekleme, silme ve değiştirme hatalarının oranı.

# Tipik İş Akışı




**Örnek:** Bir sesli asistan "yarın hava nasıl" sorusunu 16 kHz'te kaydeder, 25 ms'lik çerçevelerden spektrogram çıkarır, akustik model her çerçeve için ses birimi olasılıkları verir ve dil modeli bunları en olası cümleye dönüştürür. Metin, anlam için NLP'ye aktarılır.

DERS

# Derste Neler İşlenir?

1. Konuşmanın üretimi ve algısı; ses birimleri
2. Sayısal ses: örnekleme, çerçeveleme, pencereleme
3. Spektrogram, mel ölçeği ve MFCC
4. Saklı Markov modelleri ve Gauss karışımları
5. Viterbi ve ışın aramasıyla kod çözme
6. N-gram ve sinirsel dil modelleri
7. Derin akustik modeller
8. Uçtan uca tanıma: CTC, dikkat, dönüştürücüler
9. Değerlendirme: sözcük hata oranı
10. Konuşmacı tanıma ve doğrulama
11. Gürültüye dayanıklılık ve konuşma iyileştirme
12. Konuşma sentezi (TTS)

## Kazandırdığı beceriler

librosa ile öznitelik çıkarmak · Whisper ile döküm ve ince ayar · torchaudio ile model eğitmek · Praat ile ses çözümlemek · sistemleri WER ile karşılaştırmak

Tipik bir dönemin akışıdır; sıra ve kapsam üniversiteden üniversiteye değişir.

DENE

# Akıllı Ev Komutu Tanıyor

Akıllı ev asistanı dört sesli komutu tanımalı: evet, hayır, dur, başla. Bir komut seçin, konuşmacıyı değiştirin, gürültüyü artırın; sistem sesi kayıtlı şablonlarla karşılaştırır.

Dalga biçimi (0,6 s)



Spektrogram: 8 frekans bandının enerjisi (Hz)




**evet** | hayır | dur | başla

Başka konuşmacı | Ortam gürültüsü: **2**

Söylenen: “evet” (SNR 13 dB)
Tanınan: **evet** · güven %98 ✓

Şablonlara DTW uzaklığı (küçük = benzer)

| | |
|---|---|
| evet | 0,96 |
| hayır | 1,75 |
| dur | 2,07 |
| başla | 2,23 |

Ses önce **spektrograma**, yani zaman–frekans resmine çevrilir. Aynı sözcük hızlı ya da yavaş söylenebilir; DTW zaman eksenini esneterek eşleştirir. Gürültü arttıkça güven düşer.

KONUM

# Ön Koşullar

Bu dersten önce bilinmesi yararlı olan dersler ve neden gerektikleri.

| Ders | Neden gerekir |
|---|---|
| Sayısal Sinyal İşleme | spektrogram, MFCC, süzgeçler |
| Analiz (Kalkülüs) | karmaşık sayılar, integral ve Fourier serilerine giriş |
| Programlama | Python ile dizi işlemleri ve dosyadan ses okuma |

KONUM

# Bu Ders Sana Göre mi?

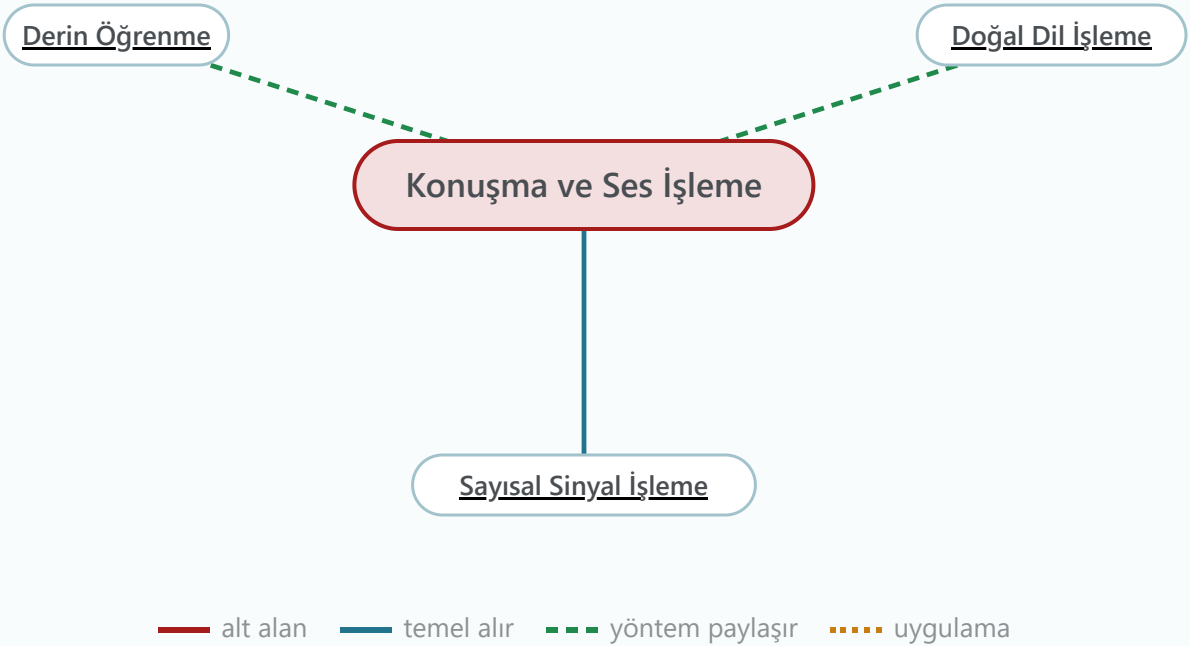
Kendine sor; "hayır" dediklerin için sağdaki sütun nereden başlayacağını gösterir.

| Kendine sor | Hayırsa önce |
|---|---|
| Fourier dönüşümünün sinyali frekanslara nasıl ayırdığını açıklayabiliyor musun? | Sayısal Sinyal İşleme |
| Koşullu olasılığı ve Bayes kuralını kullanabiliyor musun? | Olasılık ve İstatistik |
| Basit bir sinir ağının nasıl eğitildiğini açıklayabiliyor musun? | Yapay Sinir Ağları |
| Python ile bir ses dosyasını okuyup dalga biçimini çizebiliyor musun? | Programlamaya giriş, ardından NumPy ve librosa'ya kısa bir giriş |

Çoğuna "evet" diyebiliyorsan derse hazırsın; bir ya da iki "hayır" dönem içinde kapanabilir.

KONUM

# Dersin Komşuları



ÖRTÜŞME

# Örtüştüğü Dersler

| İlişki | Ders | Ortak konu |
|---|---|---|
| temel alır | Bu ders Sayısal Sinyal İşleme dersini temel alır | spektrogram, MFCC, süzgeçler |
| yöntem paylaşır | Derin Öğrenme ile yöntem paylaşır | uçtan uca konuşma tanıma |
| yöntem paylaşır | Doğal Dil İşleme ve Metin Madenciliği ile yöntem paylaşır | dil modeli, konuşmadan metne |

# Ayrıştığı Noktalar (1/2)

| Ölçüt | Konuşma ve Ses İşleme | Sayısal Sinyal İşleme | Doğal Dil İşleme |
|---|---|---|---|
| **Temel soru** | Ses dalgasından sözcükler, konuşmacı ve olaylar nasıl çıkarılır? | Örneklenmiş sinyaller nasıl çözümlenir, süzülür ve dönüştürülür? | Bilgisayar insan dilini nasıl çözümler, anlar ve üretir? |
| **Yaklaşım** | akustik sinyal ve dil modeli | matematiksel dönüşüm ve süzgeç | dilbilim, istatistik ve derin öğrenme |
| **Girdi** | ses dalgası | örneklenmiş sinyal | metin |
| **Çıktı** | metin, konuşmacı, ses olayı | süzülmüş sinyal, spektrum | etiket, çeviri, özet, yanıt |
| **Başarı ölçütü** | sözcük hata oranı (WER) | sinyal-gürültü oranı, frekans yanıtı | doğruluk, BLEU, insan değerlendirmesi |
| **Matematik yükü** | orta-yüksek | yüksek: karmaşık sayı, Fourier | orta |
| **Veri ihtiyacı** | çok saatlik kayıt | öğrenme yok; sinyal yeter | çok metin (derlem) |

- **Sayısal Sinyal İşleme:** Sinyal işleme her türlü sinyali süzer ve dönüştürür; bu ders ses ve konuşmaya özgü öznitelik ve modellere odaklanır.
- **Doğal Dil İşleme:** NLP elinde metin olduğunu varsayar; bu ders o metni ses dalgasından elde eder ve ses olaylarını da işler.

# Ayrıştığı Noktalar (2/2)

| Ölçüt | Konuşma ve Ses İşleme | Derin Öğrenme | Biyometri |
|---|---|---|---|
| **Temel soru** | Ses dalgasından sözcükler, konuşmacı ve olaylar nasıl çıkarılır? | Çok katmanlı ağlar ham veriden kullanışlı temsilleri nasıl kendisi öğrenir? | Bir kişi yüzünden, parmak izinden ya da sesinden nasıl tanınır? |
| **Yaklaşım** | akustik sinyal ve dil modeli | çok katmanlı temsil öğrenme | eşleme ve doğrulama |
| **Girdi** | ses dalgası | ham veri: görüntü, metin, ses | yüz, parmak izi, iris, ses |
| **Çıktı** | metin, konuşmacı, ses olayı | öğrenilmiş temsil ve tahmin | kimlik kararı |
| **Başarı ölçütü** | sözcük hata oranı (WER) | doğruluk, kayıp, ölçeklenme | FAR, FRR, EER |
| **Matematik yükü** | orta-yüksek | yüksek: türev, matris, olasılık | orta |
| **Veri ihtiyacı** | çok saatlik kayıt | çok; GPU gerekir | kişi başına az örnek |

- **Derin Öğrenme:** Derin öğrenme genel ağ mimarileri sunar; bu ders onları sesin zaman-frekans yapısına ve hizalama sorununa uyarlar.
- **Biyometri:** Biyometri sesi yüz ve parmak izi gibi bir kimlik kanıtı olarak kullanır; bu ders sesin içeriğini de çözümler.

UYGULAMA

# Uygulama Alanları

**Sesli asistanlar**
Komutları anlayıp telefonu ya da evi yönetmek.

**Altyazı ve döküm**
Video, toplantı ve ders kayıtlarını metne dökmek.

**Çağrı merkezi**
Görüşmeleri çözümlemek, yönlendirmek.

**Konuşmacı doğrulama**
Sesle bankacılık ve erişim denetimi.

**Ses olayı tespiti**
Cam kırılması, siren, öksürük algılama.

**Konuşma sentezi**
Metni doğal sesli konuşmaya çevirmek.

UYGULAMA

# Riskler ve Etik

**Ses klonlama**

Birkaç saniyelik kayıttan taklit edilen ses, dolandırıcılıkta ve sesle kimlik doğrulamayı aşmakta kullanılır.

**Sürekli dinleme**

Uyandırma sözcüğü bekleyen cihazlar yanlışlıkla kayıt alabilir; ses ayrıca sağlık ve duygu bilgisi taşır.

**Aksan ve ağız farkı**

Hata oranı bazı aksan ve ağızlarda belirgin yüksektir; bu konuşanlar altyazı ve asistanlardan eşit yararlanamaz.

**Uydurma döküm**

Uçtan uca modeller sessizlikte ya da gürültüde hiç söylenmemiş cümleler yazabilir; tıbbi dökümde bu tehlikelidir.

Ayrıntılı tartışma: Açıklanabilir Yapay Zekâ ve Yapay Zekâ Etiği.

UYGULAMA

# Araçlar ve Örnek Veriler

## Araçlar ve kütüphaneler

librosa | Kaldi | Praat | torchaudio | ESPnet | SpeechBrain | Whisper

## Tanınmış veri setleri

TIMIT | LibriSpeech | Mozilla Common Voice | Switchboard | VoxCeleb | AudioSet

UYGULAMA

# Sık Karıştırılanlar

| Yaygın düşünce | Doğrusu |
|---|---|
| Konuşma tanıma sesi doğrudan sözcüğe eşler. | Aynı sözcük farklı ağız, hız ve gürültüde çok farklı ses verir; dil modeli belirsizliği çözer. |
| Kayıt ne kadar yüksek sesliyse o kadar iyi. | Kırpılma (clipping) ve yankı tanımayı bozar; temiz ve dengeli kayıt önemlidir. |
| İngilizcede iyi çalışan model Türkçede de çalışır. | Her dil için yeterli veri gerekir; eklemeli Türkçede sözcük dağarcığı sorunu ayrıca çözülür. |
| Ses işleme yalnız konuşmadır. | Müzik, çevre sesleri ve tıbbi sesler (kalp, öksürük) de dersin konusudur. |

SONRASI

# Bu Dersten Sonra

## Bu dersi temel alan ya da genişleten dersler

Bu ders çoğunlukla uç noktadır; öteki dersleri temel alır.

## İlgili meslekler

Konuşma tanıma mühendisi | Ses işleme mühendisi | Araştırmacı | Gömülü ses yazılımcısı

Bu dersten sonra metnin anlamı için doğal dil işlemeye, cihaz üzerinde çalışan sesli asistanlar için gömülü YZ'ye ya da sesle kimlik doğrulama için biyometrik sistemlere geçilebilir. Sayısal sinyal işleme bilgisi ise konunun matematiğini derinleştirir.

SONRASI

# Kaynaklar


- Rabiner ve Juang, *Fundamentals of Speech Recognition*, Prentice Hall, 1993.
- Huang, Acero ve Hon, *Spoken Language Processing*, Prentice Hall, 2001.
- Jurafsky ve Martin, *Speech and Language Processing*, 2. baskı, Prentice Hall, 2009.
- Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", *Proceedings of the IEEE* 77(2), 1989.

# Son

## Bölüm 19: Konuşma Tanıma ve Ses İşleme

← Önceki bölüm · Alanın haritası · Sonraki bölüm →