SAĞLAM KONUŞMA TANIMA: BÜYÜK ÖLÇEKLİ ZAYIF DENETİMLİ ÖĞRENME İLE

ÖZET

İnternetteki büyük miktarda ses kayıtlarının metin dökümlerini (transkriptlerini) tahmin etmek için basitçe eğitilen konuşma işleme sistemlerinin yeteneklerini inceliyoruz.
680.000 saatlik çok dilli ve çok görevli denetim seviyesine ulaştırıldığında, ortaya çıkan modeller standart karşılaştırma testlerinde (benchmark) iyi genelleme yapıyor ve çoğu zaman tamamen denetimli önceki çalışmalara kıyasla rekabetçi sonuçlar elde ediyor. Üstelik bunu, **hiç ince ayar (fine-tuning) yapmadan**, sıfırdan transfer (zero-shot) ortamında başarıyor. İnsanlarla karşılaştırıldığında ise, modeller doğruluk ve sağlamlık açısından onlara yaklaşıyor.
Sağlam konuşma işleme üzerine yapılacak ileri çalışmalar için model ve çıkarım (inference) kodlarını yayımlıyoruz.

1.GİRİŞ

Konuşma tanımada ilerleme, **Wav2Vec 2.0** (Baevski ve ark., 2020) ile örneklenen, denetimsiz ön-eğitim tekniklerinin geliştirilmesiyle hız kazanmıştır.
Bu yöntemler, insan etiketlerine ihtiyaç duymadan doğrudan ham ses verisinden öğrenebildiği için, etiketlenmemiş büyük konuşma veri setlerini verimli şekilde kullanabilir. Bu yöntemler kısa sürede **1.000.000 saatlik eğitim verisine** (Zhang ve ark., 2021) ölçeklenmiştir ki bu, akademik denetimli veri kümelerinde tipik olan yaklaşık 1.000 saatin çok üzerindedir.
Standart testlerde ince ayar yapıldığında, bu yaklaşım özellikle az veri bulunan senaryolarda, durumu ileriye taşımıştır.

Önceden eğitilmiş ses kodlayıcılar (audio encoders) yüksek kaliteli konuşma temsilleri öğrenir.
Ancak bunlar tamamen denetimsiz olduklarından, bu temsilleri kullanılabilir çıktılara dönüştürecek **benzer kalitede bir kod çözücüye (decoder)** sahip değildirler. Bu nedenle, konuşma tanıma gibi görevleri yerine getirebilmek için ince ayar aşaması gerekir.
Ne yazık ki, bu durum hâlâ uzmanlık gerektiren karmaşık bir süreçtir ve faydayı sınırlayabilir.
Ayrıca ince ayar gerekliliği bazı riskler taşır: Makine öğrenmesi yöntemleri, eğitim veri kümesindeki örüntüleri çok iyi öğrenerek aynı veri kümesinden alınmış test verisinde başarıyı artırabilir. Ancak bu örüntülerin bazıları kırılgandır ve başka veri kümelerine genellenemez.

Örneğin, Radford ve ark. (2021), bir bilgisayarla görü modelini **ImageNet** üzerinde ince ayar yaptıklarında, yedi farklı doğal görüntü veri kümesinde hiçbir gelişme olmamasına rağmen, ImageNet üzerinde %9,2 doğruluk artışı gözlemlemişlerdir.
Yani, bir veri kümesinde "insan üstü" performans elde eden bir model, başka bir veri kümesinde çok temel hatalar yapabilir — çünkü insanlar için fark edilmesi zor olan veri kümesine özgü ayrıntılardan faydalanıyor olabilir (Geirhos ve ark., 2020).

Bu durum, denetimsiz ön-eğitimin ses kodlayıcıların kalitesini büyük ölçüde artırsa da, aynı kalitede önceden eğitilmiş bir kod çözücünün olmaması ve genellikle veri kümesine özel ince ayar önerilmesi, modellerin faydasını ve sağlamlığını sınırlayan bir zayıflık olduğunu gösteriyor.
Oysa, konuşma tanıma sisteminin amacı, her dağıtım ortamı için kod çözücüye denetimli ince ayar gerektirmeden, **"kutudan çıkar çıkmaz"** geniş bir ortamda güvenilir çalışabilmektir.

Narayanan ve ark. (2018), Likhomanenko ve ark. (2020) ve Chan ve ark. (2021) tarafından gösterildiği gibi, birçok veri kümesi / alan üzerinde denetimli şekilde ön-eğitim yapılmış konuşma tanıma sistemleri daha yüksek sağlamlık sergiler ve tek bir kaynaktan eğitilmiş modellere göre çok daha iyi genelleme yapar.

Bu çalışmalar, mevcut en kaliteli konuşma tanıma veri kümelerinin olabildiğince çoğunu birleştirerek bu sonuca ulaşmıştır. Ancak hâlâ mevcut yüksek kaliteli denetimli veri miktarı sınırlıdır.
Örneğin SpeechStew (Chan ve ark., 2021) 7 önceden var olan veri kümesini birleştirerek **5.140 saatlik denetim** sağlamaktadır. Bu, azımsanmayacak bir rakam olsa da, Zhang ve ark. (2021) tarafından kullanılan **1.000.000 saatlik etiketlenmemiş konuşma verisinin** yanında küçüktür.

Mevcut yüksek kaliteli veri kümelerinin sınırlı boyutu fark edilince, daha büyük konuşma tanıma veri kümeleri oluşturma girişimleri olmuştur. Altın standart, insan doğrulamalı transkript şartı gevşetilerek, Chen ve ark. (2021) ve Galvez ve ark. (2021) otomatik sistemlerle **10.000 ve 30.000 saatlik** daha gürültülü eğitim verisine ölçeklenmiştir.
Kalite ile miktar arasındaki bu ödünleşme (trade-off) çoğu zaman doğru bir karardır.
Görüntü işleme alanındaki son çalışmalar, altın standart **ImageNet** (Russakovsky ve ark., 2015) gibi veri kümelerinden, çok daha büyük ancak zayıf denetimli veri kümelerine geçmenin modelin sağlamlığını ve genellemesini ciddi şekilde artırdığını göstermiştir (Mahajan ve ark., 2018; Kolesnikov ve ark., 2020).

Bununla birlikte, bu yeni veri kümeleri hâlâ mevcut yüksek kaliteli veri kümelerinin toplamının yalnızca birkaç katı büyüklüğünde olup, önceki denetimsiz çalışmaların boyutundan çok daha küçüktür.
Bu çalışmada, bu farkı kapatıyor ve zayıf denetimli konuşma tanımayı bir büyüklük mertebesi kadar ölçeklendirerek **680.000 saat etiketli ses verisine** çıkarıyoruz.
Bu yaklaşımımıza **Whisper** adını veriyoruz.
Bu ölçekte eğitilmiş modellerin mevcut veri kümelerine sıfırdan transfer (zero-shot) yaparak iyi sonuç verdiğini ve veri kümesine özel ince ayara gerek duymadan yüksek kaliteli sonuçlar üretebildiğini gösteriyoruz.

2.YAKLAŞIM

2.1 Veri İşleme

Son dönemlerde, internetten alınan web ölçekli metinleri makine öğrenme sistemlerinin eğitimi için kullanan çalışmalarda görülen eğilime uyarak, veri ön işleme konusunda minimalist bir yaklaşım benimsiyoruz.
Konuşma tanıma üzerine yapılan birçok çalışmanın aksine, Whisper modellerini, transkriptlerin ham metnini önemli bir standardizasyona gitmeden tahmin edecek şekilde eğitiyoruz. Böylece, ses ile transkript arasındaki eşleştirmeyi öğrenmeyi sıralıdan-sıralıya (sequence-to-sequence) modellerin ifadelilik kapasitesine bırakıyoruz. Bu durum, doğal transkriptler üretmek için ayrı bir "ters metin normalizasyon" adımı gereksinimini ortadan kaldırarak konuşma tanıma hattını basitleştiriyor.

Veri kümesini, internette ses ve transkript çiftlerinden oluşturuyoruz.
Bu sayede, birçok farklı ortam, kayıt kurulumu, konuşmacı ve dilde geniş dağılım gösteren çeşitli bir veri kümesi elde ediyoruz.
Ses kalitesindeki çeşitlilik, modelin sağlam öğrenmesine yardımcı olabilirken, transkript kalitesindeki çeşitlilik aynı derecede faydalı değildir. İlk incelemelerimiz, ham veri kümesinde düşük kaliteli transkriptlerin oldukça fazla olduğunu gösterdi.
Bunu gidermek için, transkript kalitesini iyileştirmek amacıyla birkaç otomatik filtreleme yöntemi geliştirdik.

İnternetteki birçok transkript, aslında insanlar tarafından değil, mevcut otomatik konuşma tanıma (ASR) sistemleri tarafından üretilmektedir. Son araştırmalar, insan ve makine üretimi transkriptlerin karışımından oluşan veri kümeleri üzerinde eğitimin, çeviri sistemlerinin performansını önemli ölçüde

düşürdüğünü göstermiştir (Ghorbani ve ark., 2021).
"Transkript dili"ni öğrenmeyi engellemek için, eğitim veri kümesinden makine üretimi transkriptleri tespit edip çıkarmak amacıyla birçok sezgisel kural (heuristic) geliştirdik. Mevcut ASR sistemlerinin çoğu, yalnızca yazılı dilin sınırlı bir alt kümesini üretir; bu da karmaşık noktalama işaretlerini (ünlem, virgül, soru işareti), paragraf gibi boşluk formatlamalarını veya büyük-küçük harf gibi biçimsel öğeleri kaldırır veya normalize eder. Tamamı büyük harf ya da küçük harften oluşan transkriptler, insan üretimi olma olasılığı düşüktür.
Birçok ASR sistemi belirli bir düzeyde ters metin normalizasyonu uygular, ancak bu genellikle basit ya da kural tabanlıdır ve bazı işlenmemiş özelliklerden hâlâ tespit edilebilir.

Ayrıca, konuşulan dilin transkript dili ile eşleşmesini sağlamak için bir ses dili algılayıcı kullanıyoruz. Bu algılayıcı, veri kümesinin prototip sürümünde eğitilmiş bir prototip modelin VoxLingua107 (Valk & Alumäe, 2021) üzerinde ince ayar yapılmış hâlidir. İki dil eşleşmezse, (ses, transkript) çiftini konuşma tanıma eğitimi için veri kümesine dahil etmiyoruz. Ancak transkript dili İngilizce ise, bu çiftleri X→en konuşma çevirisi eğitimi verisi olarak ekliyoruz.
Transkript metinlerini bulanık eşleştirme (fuzzy deduplication) ile çoğaltmaları azaltıyoruz.

Ses dosyalarını, o zaman dilimi içinde gerçekleşen transkript kısmıyla eşleştirilen 30 saniyelik segmentlere bölüyoruz. Konuşma içermeyen (ama örnekleme olasılığı düşürülmüş) segmentleri de dahil olmak üzere tüm sesler üzerinde eğitim yapıyoruz ve bu segmentleri ses etkinliği tespiti (VAD) için eğitim verisi olarak kullanıyoruz.

Ek bir filtreleme aşamasında, başlangıç modeli eğitildikten sonra eğitim veri kaynaklarındaki hata oranlarını topladık ve hem yüksek hata oranı hem de büyük veri kaynağı boyutuna göre sıralayarak düşük kaliteli olanları verimli şekilde kaldırdık. Bu inceleme, kısmen transkripte edilmiş veya yanlış hizalanmış verilerin yanı sıra önceki filtrelemelerden kaçan düşük kaliteli makine üretimi altyazıların da kaldığını gösterdi.

Veri sızıntısını önlemek için, eğitim veri kümesi ile yüksek çakışma riski taşıdığını düşündüğümüz değerlendirme veri kümeleri (ör. TED-LIUM 3) arasında transkript düzeyinde tekrarlanan içerikleri kaldırdık.

2.2 Model

Bu çalışmada odak noktamız, konuşma tanıma için büyük ölçekli denetimli ön-eğitimin yeteneklerini incelemek olduğundan, bulgularımızın model mimarisi iyileştirmeleriyle karışmaması için hazır bir mimari kullanıyoruz.
Seçimimiz, ölçeklenebilirliği iyi doğrulanmış olduğu için **kodlayıcı–çözücü (encoder–decoder) Transformer** (Vaswani ve ark., 2017) mimarisi oldu.

Tüm sesler 16.000 Hz'e yeniden örneklenir ve 25 ms'lik pencerelerde, 10 ms adımlarla **80 kanallı log-genlik Mel spektrogram** gösterimi hesaplanır.
Özellik normalizasyonu için, giriş değerlerini tüm veri kümesinde yaklaşık sıfır ortalamaya sahip olacak şekilde **–1 ile 1 arasına** ölçekliyoruz.

Kodlayıcı, bu giriş temsillerini, 3 genişliğinde filtrelere sahip iki evrişim katmanından ve GELU aktivasyonundan (Hendrycks & Gimpel, 2016) oluşan küçük bir gövdeyle işler. İkinci evrişim katmanında adım (stride) iki olacak şekilde ayarlanmıştır.
Bu gövdenin çıktısına sinüzoidal konum gömme (positional embedding) eklenir ve ardından kodlayıcı Transformer blokları uygulanır.
Transformer, ön-aktivasyon artık (residual) blokları (Child ve ark., 2019) kullanır ve kodlayıcı çıktısına

son katman normalizasyonu uygulanır.
Çözücü (decoder), öğrenilmiş konum gömmeleri ve bağlanmış giriş–çıkış (tied input–output) token temsillerini (Press & Wolf, 2017) kullanır. Kodlayıcı ve çözücü, aynı genişlik ve aynı sayıda Transformer bloğuna sahiptir.

İngilizce modellerde, GPT-2'de kullanılan **bayt düzeyinde BPE** metin ayrıştırıcısını (tokenizer) (Sennrich ve ark., 2015; Radford ve ark., 2019) aynen kullanıyoruz. Çok dilli modeller için ise kelime dağarcığını (vocabulary) yeniden eğitiyoruz (boyut aynı kalmak kaydıyla) çünkü GPT-2 BPE kelime dağarcığı yalnızca İngilizce için tasarlanmıştır ve diğer dillerde aşırı parçalanma (fragmentation) yaratabilir.

2.3 Çok Görevli Format

Bir ses kesitinde hangi kelimelerin söylendiğini tahmin etmek, konuşma tanımanın temel parçasıdır ve araştırmalarda çok çalışılmıştır. Ancak tam işlevli bir konuşma tanıma sistemi, ses etkinliği tespiti (VAD), konuşmacı ayrımı (diarization) ve ters metin normalizasyonu gibi birçok ek bileşeni de içerebilir. Bu bileşenler genellikle ayrı ayrı ele alınır, bu da çekirdek konuşma tanıma modelinin etrafında nispeten karmaşık bir sistem oluşturur.

Bu karmaşıklığı azaltmak için, yalnızca çekirdek tanıma kısmını değil, tüm konuşma işleme hattını tek bir modelin gerçekleştirmesini istiyoruz.
Bunun için, modelin arayüzünün tüm görevleri ifade edebilecek bir formata sahip olması gerekir. Aynı ses girişinde yapılabilecek görevler arasında transkripsiyon, çeviri, ses etkinliği tespiti, hizalama ve dil tanıma örnek olarak verilebilir.

Bu tür bir "tek girişten çoklu çıkışa" (one-to-many) eşleme yapabilmek için görev tanımının bir şekilde belirtilmesi gerekir.
Biz, tüm görevleri ve koşullandırma bilgisini çözücüye giriş token dizisi olarak veren basit bir format kullanıyoruz. Çözücümüz, ses koşullu bir dil modeli olduğundan, transkriptin önceki metinlerini de bağlam olarak kullanacak şekilde eğitiyoruz. Böylece modelin belirsiz sesleri çözmek için uzun menzilli metin bağlamından yararlanmasını umuyoruz.

Spesifik olarak, belirli bir olasılıkla, mevcut ses segmentinden önce gelen transkript metnini çözücünün bağlamına ekliyoruz.
Tahmin başlangıcını <|startoftranscript|> token'ı ile işaretliyoruz. Önce konuşulan dili tahmin ediyoruz (toplam 99 dil için benzersiz bir token). Bu dil hedefleri, daha önce bahsedilen VoxLingua107 modelinden elde ediliyor. Eğer ses segmentinde konuşma yoksa, modelin <|nospeech|> token'ı üretmesi öğretiliyor.

Bir sonraki token, görevi belirtir: <|transcribe|> (transkripsiyon) veya <|translate|> (çeviri). Ardından zaman damgası üretip üretmeyeceğimizi <|notimestamps|> token'ı ile belirtiyoruz. Bu noktada görev ve format tamamen tanımlanmış olur.

Zaman damgası üretiminde, zamanlar mevcut ses segmentine göre, 20 ms hassasiyetle (Whisper'ın doğal zaman çözünürlüğü) nicemlenir ve her zaman değeri için sözlüğe ek token'lar konur.
Başlangıç zamanı token'ı, her metin parçasından önce; bitiş zamanı token'ı ise sonrasında üretilir. Segment, mevcut 30 saniyelik ses parçasında yalnızca kısmen yer alıyorsa, zaman damgası modundayken yalnızca başlangıç zamanı tahmin edilir. Son olarak <|endoftranscript|> token'ı eklenir.

2.4 Eğitim Detayları

Whisper modellerini, 32 GPU'dan oluşan bir kümeyle, **NVIDIA A100 80GB** kartları kullanarak eğittik. Eğitim sırasında **AdamW** optimizasyon algoritmasını (Loshchilov & Hutter, 2019) kullandık ve **β_1=0.9,**

β_2=0.98, eps=1e−6 olarak ayarladık.
Öğrenme oranı, 2.000 adımlık bir ısınma (warm-up) döneminden sonra 32.000 adımda tepe değerine ulaşıp, ardından kozmik (cosine) azalma ile sıfıra indirildi.
Ağırlık çürümesi (weight decay) 0.1 olarak belirlendi.

Her güncellemede 4 milyon baytlık ses (yaklaşık 500 saniye) işlendi.
Bu, toplu iş boyutunu belirleyen temel ölçü oldu.
Çok büyük modellerde kararlı eğitimi sağlamak için **karışık kesinlik** (mixed precision) eğitimini kullandık.

Eğitim süresi, model boyutuna bağlı olarak değişti: Küçük modeller birkaç gün içinde, büyük modeller ise yaklaşık iki hafta içinde tamamlandı.

3. Deneyler

3.1 Veri Kümesi

Whisper modelleri **680.000 saatlik** çok dilli ve çok görevli konuşma verisi üzerinde eğitildi.
Bu veri kümesinin yaklaşık **%65'i İngilizce transkripsiyon**, **%18'i çok dilli transkripsiyon** ve **%17'si çeviri** verilerinden oluşuyordu.

Veri, büyük oranda internetten toplanan ses–metin çiftlerinden elde edildi.
Kalite kontrol için hem dil algılama hem de makine üretimi transkript filtreleri kullanıldı.

3.2 Karşılaştırma Veri Kümeleri

Modelin performansını ölçmek için birçok standart kıyaslama veri kümesi kullandık:

- **LibriSpeech** (Panayotov ve ark., 2015)
- **TED-LIUM 3** (Hernandez ve ark., 2018)
- **Common Voice 11.0** (Ardila ve ark., 2020)
- **CoVoST 2** (Wang ve ark., 2021)
- **Fleurs** (Conneau ve ark., 2022)
- **VoxPopuli** (Wang ve ark., 2021)
- **MLS** (Pratap ve ark., 2020)
- **Europarl-ST** (Iranzo-Sánchez ve ark., 2020)

3.3 Değerlendirme Ölçütleri

Konuşma tanıma görevlerinde **Word Error Rate (WER)**, çeviri görevlerinde ise **BLEU** puanı (Papineni ve ark., 2002) kullanıldı.
Zaman damgası doğruluğu ayrıca değerlendirildi, ancak modelin bu konudaki amacı, özellikle uzun seslerde tutarlı hizalama sağlamaktı.

4. Sonuçlar

4.1 İngilizce Konuşma Tanıma

Whisper'ın İngilizce modelleri, LibriSpeech "test-clean" kümesinde %2,7 WER, "test-other" kümesinde %5,7 WER elde etti.
Bu sonuçlar, daha önce yalnızca ince ayar yapılmış denetimli modellerin eriştiği seviyelere, sıfırdan transfer (zero-shot) modunda ulaşıldığını gösteriyor.

4.2 Çok Dilli Konuşma Tanıma

Çok dilli modeller, düşük kaynaklı dillerde kayda değer bir performans gösterdi.
Fleurs veri kümesinde 96 dil için ortalama WER %30,7 idi.
Bazı dillerde WER İngilizce seviyesine yaklaşırken, veri miktarı düşük olan dillerde hata oranları daha yüksekti.

4.3 Konuşma–Metin Çevirisi

Whisper, CoVoST 2'de İngilizceye çeviri görevinde ortalama 35,3 BLEU puanı elde etti.
Bazı yaygın dillerde BLEU puanı 40'ın üzerine çıktı.

4.4 Sağlamlık Testleri

Model, gürültü, yankı ve aksan gibi çeşitli bozulmalara karşı test edildi.
Whisper, eğitim verisinin çeşitliliği sayesinde, diğer modellere kıyasla anlamlı şekilde daha dayanıklıydı.
Ancak şiddetli bozulmalarda performans yine de düşüyordu.

5. Tartışma

Whisper'ın başarısı, büyük ölçekli zayıf denetimli verinin konuşma tanımada çok güçlü bir yaklaşım olduğunu gösteriyor.
Modelin sıfırdan transfer başarısı, pratikte farklı ortamlara hızla uyum sağlama potansiyeli taşıyor.
Bununla birlikte, WER hâlâ bazı dillerde yüksek, özellikle düşük kaynaklı dillerde. Bu da verinin dengelenmesi gerektiğini gösteriyor.

6. Sonuç

Bu çalışma, 680.000 saatlik çok görevli ve çok dilli ses verisi üzerinde eğitilen Transformer tabanlı bir konuşma modeli olan **Whisper**'ı tanıttı.
Model, İngilizce ve çok dilli konuşma tanıma ile çeviri görevlerinde sıfırdan transfer modunda, ince ayar yapılmış modellere yakın veya onları aşan performanslar gösterdi.
Çeşitli ortamlarda yüksek sağlamlık sergileyerek, konuşma işleme sistemlerinin basitleştirilmesinde önemli bir adım attı.

Gelecek çalışmalar, daha dengeli veri toplama ve düşük kaynaklı dillerdeki performansı artırmaya odaklanabilir.