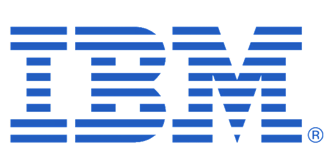
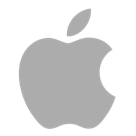
# Speech To Text

Otomatik Konuşma Tanıma (Automatic Speech Recognition - ASR) olarak da adlandırılan konuşmadan metne (Speech-to-text - STT) sanal asistanlar, otomatik altyazı oluşturma, not alma ve daha fazlası gibi uygulamaları besleyen, hızla büyüyen bir teknolojidir.

STT'yi kendi ürünlerinize entegre etmek için üç ana seçenek vardır: büyük teknoloji şirketlerinden bulut hizmetleri, uzman ses teknolojisi şirketleri ve kendi teknolojinizi şirket içinde oluşturmak. Hangisinin en iyi olduğuna karar vermek, genellikle bunları, ürününüzde görmeyi beklediğiniz türden temsil eden bazı veriler üzerinde test etmek ve farklı teklifleri kendi bütçeniz ve gereksinimlerinizle eşleştirmek anlamına gelir.

**Büyük Bulut Teknoloji Şirketleri**

Büyük teknoloji şirketlerinin çoğu, bulut teklifleri paketinin bir parçası olarak STT hizmetlerine sahiptir. Amazon, Google, Microsoft, IBM ve Baidu'nun API'ler olarak sunulan hizmetleri vardır ve Apple'ın STT'si geliştirici programı aracılığıyla kullanılabilir. Bu API'ler, gerçek zamanlı akış modunda çalışmak veya toplu ses yüklemek için uygulamalara ve diğer ürünlere entegre edilebilir

Uygulamanıza özel kelimeleri işlemek için özel kelime listeleri ve telaffuzlar gibi bazı sınırlı özelleştirme seçenekleri mevcuttur. Farklı diller mevcuttur, ancak API'nin arkasında çalışan model hakkında genellikle birkaç seçenek vardır. Bu hizmetlerin iyi tarafı, entegre edilmelerinin kolay olması ve genel kullanım durumlarında performanslarının genellikle iyi olmasıdır. İnşa etmek için zaman ve çaba harcamanız gerekmediğinden, sadece kredi kartıyla bir hesap açın ve kendiniz deneyin, başlamanın hızlı ve kolay bir yolu olabilirler. Fiyatlar halka açıktır. Google'ın veri kaydı olmayan standart modelleri için fiyat, 15 saniyelik ses başına 0,6 senttir. Amazon'un ayda 250 bin dakikaya kadar yazıya dönüştürme fiyatı dakikada 2,4 sent ve daha yüksek hacimler için daha ucuz. Bu fiyatlardan bulut STT hizmetlerinin çağrı merkezleri veya otomatik altyazı gibi yüksek hacimli uygulamaları hedef aldığını görebilirsiniz.

**Specialist Ses Şirketleri**

Büyük teknoloji şirketlerinin ötesinde, bir dizi farklı STT ürünü sunan bir dizi küçük ve orta ölçekli işletme var. Bunların hepsinin farklı fiyatlandırmaları, satış noktaları ve özellikleri var, bu yüzden kısaca özetlemek zor. Hangisinin hedeflerinize uyduğunu görmek için onları daha derinlemesine incelemeye değer. Bazı örnekler:

* Otter — toplantılar, röportajlar ve diğer konuşmalar için zengin notlar
* Cobalt — özel, ısmarlama ses teknolojisi çözümleri
* Speechmatics — doğru ve kapsamlı konuşma tanıma motoru
* Picovoice edge cihazlarda ses teknolojisi
* Rev/Temi endüstriler arasında hem manuel hem de otomatik transkripsiyon
* Verbit yasal ve eğitim ürünlerine uyarlanmıştır
* Deepgram — özel ASR modelleme
* Descript — içerik oluşturucular için ses ve video düzenleme
* SoundHound — özel sesli yardımcılar ve teknoloji oluşturma
* Soapbox Labs — özellikle çocuklar için konuşma teknolojisi geliştiriyor

**Şirket İçi Geliştirme**

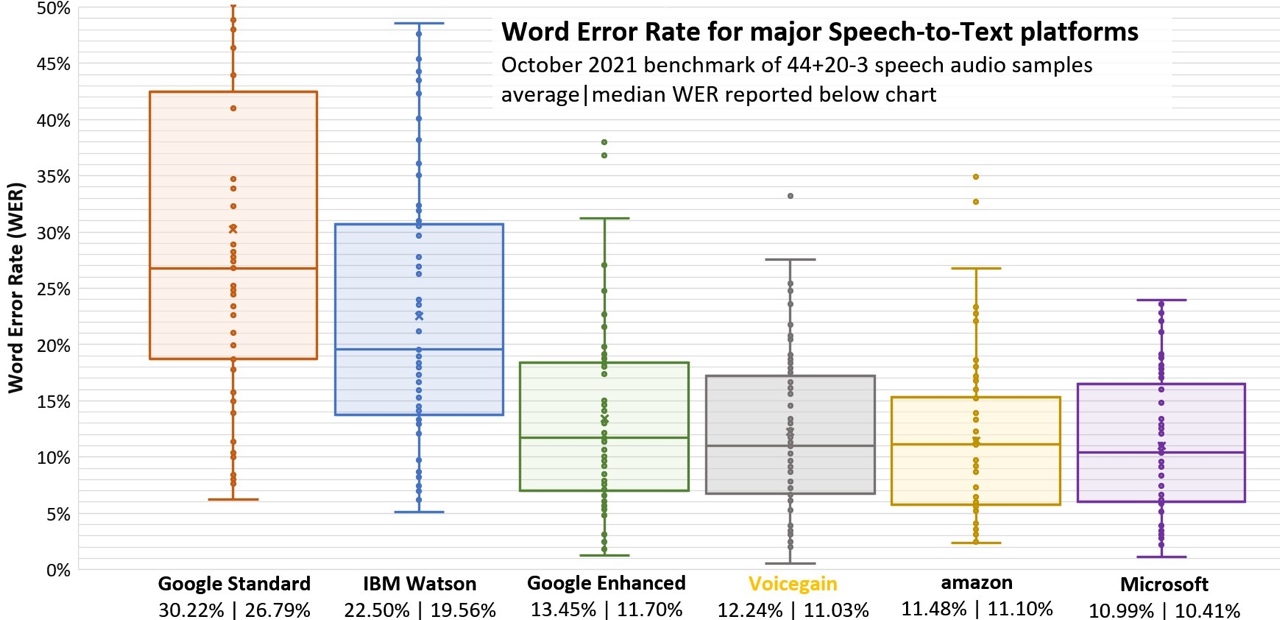
Üçüncü ve en çok zaman alan seçenekler, kendi teknolojinizi kurum içinde oluşturmaktır. Mevcut çözümler aradığınızı sunmuyorsa, bu seçeneği takip edebilirsiniz. Belki de, bulut API'lerinin arkasındaki genel amaçlı modellerin yeterince doğru olmadığı belirli bir alan için veya bulut bağlantısına güvenemeyeceğiniz bir cihaz için, bunlar tarafından iyi hizmet verilmeyen bir dil için geliştiriyorsunuz.

Bir konuşma tanıma sistemi oluşturmak için ihtiyaç duyacağınız bileşenler, onu eğitmek için bol miktarda veri, eğitim için bilgi işlem gücü ve modelleri oluşturup ayarlamak için bir veya daha fazla makine öğrenimi uzmanıdır. Bu becerilerin işe alınmasının zor olduğunu ve makine öğrenimi uzmanlarına yüksek talep geldiğinden maaşların genellikle yüksek olduğunu belirtmekte fayda var.

Başlamak ve tamamen sıfırdan başlama işini kurtarmak için bir dizi açık kaynaklı konuşma araç takımı vardır. **Kaldi**, bunların en bilinenlerinden ve kullanılanlarından biridir, ancak ses teknolojisi alanında yeni olanlar için dik bir öğrenme eğrisi vardır. Daha yakın zamanda duyurulan bir diğeri ise **SpeechBrain**. Başka hangi kaynakların mevcut olduğunu öğrenmek için **OpenSLR'**nin iyi bir listesi var.

Kendi teknolojinizi oluşturmak ve dağıtmak için haftalar veya aylar harcamayı bekleyebilirsiniz ve ardından onu sürdürmek için yatırım yapmanız gerekecektir. Ancak sonuç, daha esnek, gereksinimlerinize göre ayarlanmış ve ürününüze derinlemesine entegre edilmiş bir sistem olacaktır.

# Speech-to-Text Doğruluk Kıyaslaması



**"Tanıyıcınızın doğruluğuna neler etki eder ?"**

Genellikle "buna bağlıdır" cevabını veririz ve diğer tarafın "doğru bir cevap vermezlerse gerçekten kötü olması gerektiğini" düşündüğü hissine kapılırız. Ancak, "buna bağlı" gerçekten doğru cevaptır. Otomatik konuşma tanımanın (ASR) doğruluğu birçok yönden sese bağlıdır ve etkisi küçük değildir. Temel olarak, aşağıdaki gibi faktörlere bağlı olarak doğruluk her yerde olabilir:

* Konuşma uygun dilbilgisini takip ediyor mu yoksa konuşmacı her şeyi söylediği gibi mi uyduruyor? Hazırlanmış konuşmalar, yazılmamış konuşmaya kıyasla daha iyi, yani daha düşük WER (word error rate - kelime hata oranı) puanlarına sahip olacaktır.
* Konuşmanın konusu nedir. Nadir ve anlaşılması güç kelimeler veya kelime kombinasyonları, ör. kişiler veya diğer isimler, NLM (Natural Language Model - doğal dil modeli) için hayatı zorlaştıracaktır.
* Birden fazla konuşmacı var mı? Sürekli geçiş yapıyorlar mı, hatta birbirleri hakkında konuşuyorlar mı?
* Arka planda müzik var mı - youtube yapımları için çok yaygın.
* Arka plan gürültüsü var mı? Gürültünün türü nedir?
* Konuşma sesinin bazı bölümleri alışılmadık şekilde yavaş mı yoksa hızlı mı?
* Kayıtta oda yankısı veya yankısı var mı?
* Kayıt hacmi çok mu düşük? Kayıt hacminde farklılıklar var mı (örneğin, çok uzun bir masanın bir kenarına yerleştirilmiş kayıt cihazı)
* Örneğin, bir codec bileşeni veya çılgın arşiv sıkıştırma seviyeleri nedeniyle kayıt kalitesi kötü mü?

**Konuşmadan Metne Doğruluğu Test Etme / Karşılaştırma**

Doğruluk veya Kelime Hata Oranı soruları, konuşma sesinin türünü belirtmeden biraz anlamsız olduğundan, bir konuşma tanıyıcı seçerken test yapmak önemlidir. Bir test seti olarak, beklenen kullanım durumlarında tanıyıcı tarafından karşılaşılacak konuşma spektrumunu doğru bir şekilde temsil eden bir dizi ses dosyası seçilecektir. Setteki her konuşma ses dosyası için %100 doğru bir altın/referans dökümü elde edilir. Bundan sonra, işler otomatik hale getirilebilir - değerlendirilmekte olan tanıyıcılardaki her dosyayı kopyalayın, oluşturulan transkriptlerin her biri için referansa karşı WER'i hesaplayın ve sonuçları harmanlayın. Birleştirilmiş sonuçlar, tanıyıcıların önemsediğimiz belirli konuşma sesinde nasıl performans gösterdiğinin net bir resmini sunacaktır. Bu işlemi sık sık tekrarlayacaksanız, örneğin tanıyıcı işaretçideki yeni adayları değerlendirmek için, temel olarak gelecekte referans alınabilecek tekrarlanabilir bir kıyaslama oluşturarak test setini standart hale getirmek iyidir.

**Sonuçlar neler? En iyi tanıyıcı kimde?**

En iyi tanıyıcı doğru soru değildir, çünkü bunların tümü, kullanıldığı gerçek konuşma sesinize bağlıdır. Yapılan testlerden elde edilen önemli sonuçlar aşağıdaki gibidir:

* Her tanıyıcı geliştirildi. Medyan WER'deki en büyük gelişme Microsoft Speech to Text tarafından yapıldı.
* Veri kümesindeki en iyi tanıyıcı Google Speech to Text - Enhanced (video) idi, ancak yeni Microsoft Speech to Text çok yakın ikinci sırada.
* Fiyat dikkate alındığında, Microsoft Best Buy ilan edilebilir
* Google Speech to Text - Standart, biraz geliştirilmiş olmasına rağmen, veri setinde hala en kötü performans gösterendir.

**Açık Kaynaklı Speech to Text Araçları**

1. **Project DeepSpeech:**   
   Bu proje, Firefox tarayıcısının arkasındaki kuruluş olan Mozilla tarafından yapılmıştır. Misyonunu yerine getirmek için TensorFlow çerçevesini kullanan makine öğrenimi teknolojisini de içeren %100 ücretsiz ve açık kaynaklı bir konuşmadan metne kitaplıktır. Başka bir deyişle, temeldeki konuşmadan metne teknolojisini geliştirmek ve daha iyi sonuçlar almak için eğitim modellerini kendiniz oluşturmak ve hatta isterseniz diğer dillere getirmek için kullanabilirsiniz. Ayrıca, TensorFlow'da sahip olduğunuz diğer makine öğrenimi projelerinize de kolayca entegre edebilirsiniz. Ne yazık ki, proje şu anda varsayılan olarak yalnızca İngilizce'yi destekliyor gibi görünüyor. Python (3.6) gibi birçok dilde de mevcuttur. Ancak, son Mozilla yeniden yapılanmasından sonra, neye karar vereceklerine bağlı olarak kapatılabileceği (veya kapatılmayacağı) için projenin geleceği bilinmiyor.  
   **Repo Url:** [**https://github.com/mozilla/DeepSpeech**](https://github.com/mozilla/DeepSpeech)
2. **Kaldi**  
   Kaldi, C++ ile yazılmış açık kaynaklı bir konuşma tanıma yazılımıdır ve Apache kamu lisansı altında yayınlanmıştır. Windows, macOS ve Linux'ta çalışır. Gelişimi 2009'da başladı. Kaldi'nin diğer bazı konuşma tanıma yazılımlarına göre ana özelliği, genişletilebilir ve modüler olmasıdır: Topluluk, görevleriniz için kullanabileceğiniz tonlarca 3. taraf modülü sağlar. Kaldi ayrıca derin sinir ağlarını destekler ve web sitesinde mükemmel bir dokümantasyon sunar. Kod çoğunlukla C++ ile yazılırken, Bash ve Python betikleri tarafından "sarılır".  
   Bu nedenle, yalnızca konuşmayı metne dönüştürmenin temel kullanımını arıyorsanız, bunu Python veya Bash aracılığıyla gerçekleştirmenin kolay olduğunu göreceksiniz. Ayrıca, kullanıma hazır İngilizce eğitimli modelleri olan önceden oluşturulmuş bir Python motoru olan Kaldi Active Grammar'ı da kontrol etmek isteyebilirsiniz. Resmi web sitesinden Kaldi konuşma tanıma hakkında daha fazla bilgi edinin.  
   **Resmi Websitesi:** [**http://kaldi-asr.org/**](http://kaldi-asr.org/)
3. **Julius**1991'de Kyoto Üniversitesi'nde geliştirilmeye başlandığı ve daha sonra mülkiyeti 2005'te bağımsız bir proje olarak devredildiği için muhtemelen şimdiye kadarki en eski konuşma tanıma yazılımlarından biri. Birçok açık kaynak uygulaması onu motor olarak kullanıyor (Think of KDE Simon). Julius'un ana özellikleri arasında gerçek zamanlı STT işlemleri gerçekleştirme yeteneği, düşük bellek kullanımı (20000 kelime için 64 MB'den az), N-best/Kelime grafiği çıktısı üretme yeteneği, bir sunucu birimi olarak çalışabilme yeteneği ve çok daha fazlası yer alır. Bu yazılım temel olarak akademik ve araştırma amaçlı olarak oluşturulmuştur. C dilinde yazılmıştır ve Linux, Windows, macOS ve hatta Android'de (akıllı telefonlarda) çalışır. Şu anda yalnızca İngilizce ve Japonca dillerini desteklemektedir. Yazılım, muhtemelen Linux dağıtımınızın deposunu kullanarak kolayca kurulabilir; Paket yöneticinizde julius paketini arayın. Julius kaynak koduna GitHub'dan ulaşabilirsiniz.  
   **Github Kaynak Kodları:** [**https://github.com/julius-speech/julius**](https://github.com/julius-speech/julius)
4. **Wav2Letter++**Modern bir şey arıyorsanız, o zaman bu sizin için. Wav2Letter++, Facebook'un AI Araştırma Ekibi tarafından piyasaya sürülen açık kaynaklı bir konuşma tanıma yazılımıdır. Kod, BSD lisansı altında yayınlandı. Facebook, kütüphanesini "mevcut olan en hızlı son teknoloji konuşma tanıma sistemi" olarak tanımlıyor. Bu aracın üzerine inşa edildiği kavramlar, varsayılan olarak performans için optimize edilmesini sağlar; Facebook'un aynı zamanda yeni makine öğrenimi kitaplığı FlashLight, Wav2Letter++'ın temel çekirdeği olarak kullanılıyor. Wav2Letter++, algoritmaları eğitmek için önce kendi dileğiniz için bir eğitim modeli oluşturmanıza ihtiyaç duyar. Herhangi bir dilin (İngilizce dahil) önceden oluşturulmuş desteği yoktur. Konuşmayı metne dönüştürmek için yalnızca makine öğrenimine dayalı bir araçtır. C++ ile yazılmıştır, dolayısıyla adı (Wav2Letter++).  
   **Source Code**: **https://github.com/flashlight/wav2letter**
5. **DeepSpeech2**Çin merkezli dev şirket Baidu'daki araştırmacılar, DeepSpeech2 adı verilen kendi konuşmadan metne motorları üzerinde de çalışıyorlar. Hem İngilizce hem de Mandarin Çincesi konuşmalarını metne dönüştürmek için "PaddlePaddle" derin öğrenme çerçevesini kullanan uçtan uca açık kaynaklı bir motordur. Kod BSD lisansı altında yayınlandı. Motor, istediğiniz modelde ve istediğiniz dilde eğitilebilir. Modeller kodla birlikte piyasaya sürülmez. Tıpkı diğer yazılımlar gibi, bunları kendiniz oluşturmanız gerekecek. DeepSpeech2'nin kaynak kodu Python'da yazılmıştır, bu nedenle kullandığınız dil buysa, ona aşina olmanız kolay olacaktır.  
   **Source Code: https://github.com/PaddlePaddle/PaddleSpeech**
6. **OpenSeq2Seq**NVIDIA tarafından diziden diziye modeller eğitimi için geliştirilmiştir. Konuşma tanımadan çok daha fazlası için kullanılabilse de, bu kullanım durumu için yine de iyi bir motordur. Bunu kullanarak kendi eğitim modellerinizi oluşturabilir veya varsayılan olarak gönderilen Jasper, Wave2Letter+ ve DeepSpeech2 modellerini kullanabilirsiniz. CUDA ve güçlü grafik kartları gibi bazı NVIDIA teknolojileri için yoğun desteğin yanı sıra, birden çok GPU/Çoklu CPU kullanarak paralel işlemeyi destekler.
7. **Fairseq**Başka bir diziden diziye araç seti. Facebook tarafından geliştirildi ve Python ve PyTorch çerçevesinde yazılmıştır. Ayrıca paralel eğitimi destekler. Çeviri ve daha karmaşık dil işleme görevleri için bile kullanılabilir. **Source Code: https://github.com/pytorch/fairseq**
8. **Vosk**Geliştirilmesi henüz 2020'de başladığı için en yeni açık kaynaklı konuşma tanıma sistemlerinden biri. Bu listedeki diğer sistemlerden farklı olarak Vosk, 10 dili (İngilizce, Almanca, Fransızca, Türkçe…) desteklediği için kurulumdan sonra kullanıma oldukça hazırdır (Gerekirse 1.4 GB'a kadar daha büyük modeller var). Ayrıca Raspberry Pi, iOS ve android cihazlarda çalışır ve konuşma tanıma görevlerinizi çevrimiçi yapmak için ona bağlanmanıza izin veren bir akış API'sı sağlar. Vosk'un Java, Python, JavaScript, C# ve NodeJS için bağlantıları vardır.  
   **Official Website: https://alphacephei.com/vosk/**
9. **Athena**ASR'yi (Otomatik konuşma tanıma) uygulayan uçtan uca bir konuşma tanıma motoru. Python ile yazılmıştır ve Apache 2.0 lisansı ile lisanslanmıştır. Denetimsiz ön eğitim ve çoklu GPU işlemeyi destekler. TensorFlow'un üzerine inşa edilmiştir  
   **Source Code: https://github.com/athena-team/athena**
10. **ESPnet**PyTorch'un üstüne Python ile yazılmıştır. Ayrıca uçtan uca ASR'yi destekler. Veri işleme için Kaldi stilini takip eder, bu nedenle ondan ESPnet'e geçiş yapmak daha kolay olacaktır. ESPnet'in ana pazarlama noktası, birçok kıyaslamada sunduğu son teknoloji ürünü performans ve metinden konuşmaya (STT), makine çevirisi (MT) ve konuşma çevirisi (ST) gibi diğer dil işleme görevlerine verdiği destektir. .  
    **Documentation: https://espnet.github.io/espnet/**

###### Kaynaklar:

1. <https://medium.com/mlearning-ai/a-rough-guide-to-the-speech-to-text-landscape-cdde2c3eaf7e>
2. https://www.voicegain.ai/post/speech-to-text-accuracy-benchmark-june-2020-results
3. https://fosspost.org/open-source-speech-recognition/