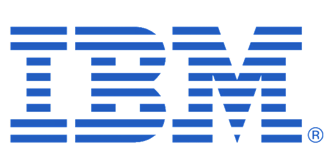
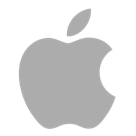
# Speech To Text

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**

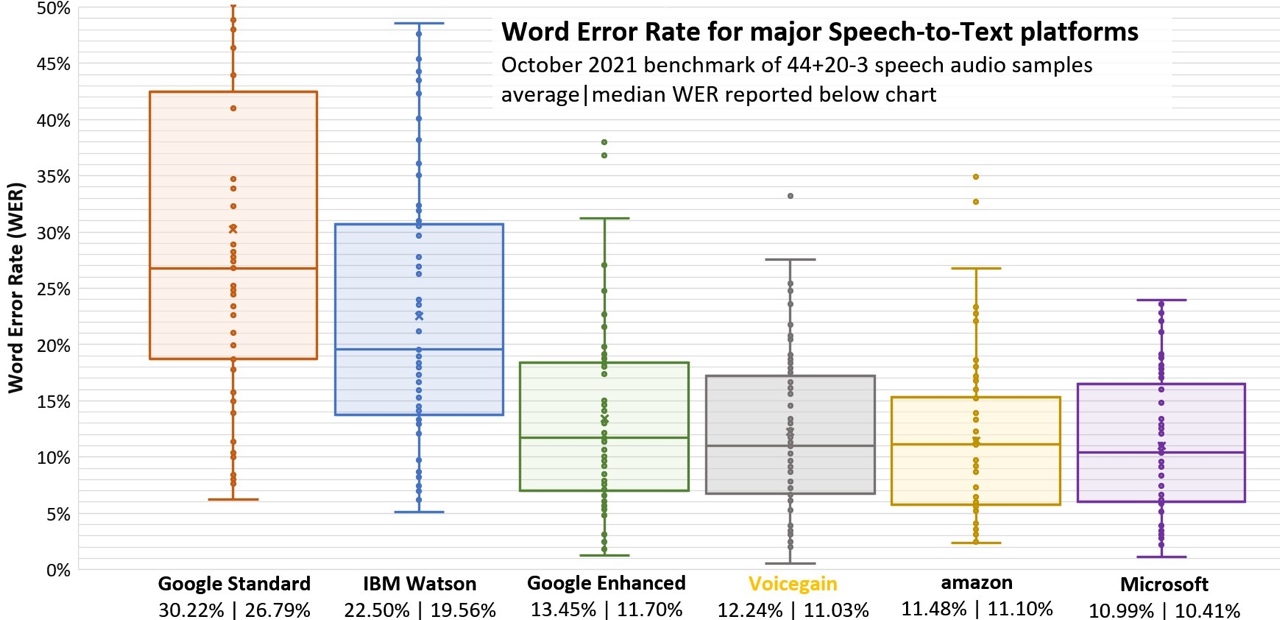
Üçüncü ve en çok zaman alan seçenekler, kendi teknolojinizi kurum içinde oluşturmaktır. Mevcut çözümler aradığınızı sunmuyorsa, bu seçeneği takip edebilirsiniz. Belki de, bulut API'lerinin arkasındaki genel amaçlı modellerin yeterince doğru olmadığı belirli bir alan için veya bulut bağlantısına güvenemeyeceğiniz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

Test

###### 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