BİLGİSAYAR DESTEKLİ BİR DİL PROGRAMI
-Türkçe Konuşma - Tanıma Sistemi-

Prof. Dr. Fatih KİRİŞÇİOĞLU
Bilgisayarlı Dil Uzmanı Erkan KARABACAK
Proje Sorumlusu Çetin ÇETİNTÜRK

Tanımlar :

Konuşma Tanıma : Bilgisayar yazılımı ile konuşmanın yazıya dönüştürülmesini ifâde eder.

Dil Modeli : Dilin söz ve ek diziliş kurallarının bilgisayar yazılımı ile modellenmesini ifâde eder.

Gerçek Zamanlı Konuşma Tanıma : Konuşmanın yazıya dönüştürülmesi işleminin süre olarak konuşmadan daha kısa olmasını ifâde eder.

Kelime Kapasitesi : Konuşma tanıma sisteminin tanıyabileceği birbirinden farklı kelime sayısıdır. Ek alarak şekil değiştiren kelimeler farklı kelime sayılmaktadır.

ÖZET

İnsan için en doğal ve en kolay iletişim yöntemi konuşmadır. 21nci yüzyılın ileri teknolojili-akıllı cihazların çağı olacağını beklemekteyiz. Akıllı olmanın ve kolay kullanılabilir olmanın temel şartlarından birisi de konuşma tanıyabilmektir. Bu çağda yaşayan insanlar olarak bilgisayardan cep telefonuna, otomobile kadar pek çok cihazı konuşarak kullanabileceğimiz bir teknolojiye oldukça yaklaşmış durumdayız.

Konuşma tanıma teknolojisi özellikle, 2000'li yıllarda yaygınlaşmaya başlayan İngilizce için uzun zamandır geliştirilmekte olan bir teknolojidir. Güzel Türkçemizin de bu konuda önde olması için özel bir yazılım firması tarafından uzun süredir bir araştırma projesi yürütülmektedir.

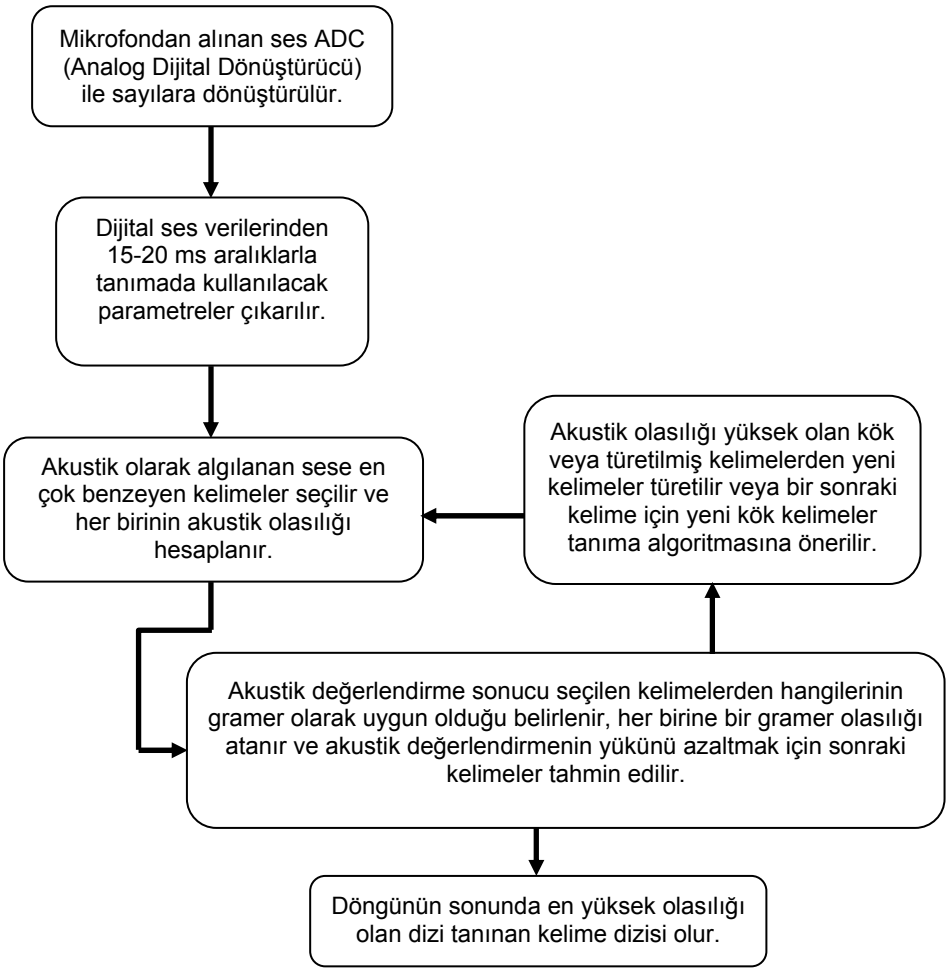
Söz konusu proje çerçevesinde Dünyada ilk defa Türkçe için gerçek zamanlı ve yüksek kapasiteli bir konuşma tanıma sistemi geliştirilmiştir. Geliştirilmiş olan teknoloji Dünya üzerindeki benzer teknolojilere kıyasla önemli üstünlüklere sahiptir.

Söz konusu sistem milyonlarca Türkçe kelimeyi (ek alarak şekil değiştiren kelime farklı kelime sayılmaktadır) gerçek zamanlı (konuşma süresinden daha kısa sürede) olarak tanıyabilmektedir. Sistemin sağlıklı olarak çalışabilmesi için Dünya'nın en karmaşık dil modellerinden biri oluşturulmuştur. Bu çalışma sırasında internetten indirilen milyarlarca kelimeden oluşan elektronik metin veri tabanı otomatik yazılımlarla analiz edilmiş ve bu analiz sonucunda yapısal bir Türkçe dil modeli oluşturulmuştur.

Geliştirilen Türkçe konuşma tanıma teknolojisi ülkemiz ve Türkçe konuşan diğer ülkeler için çok önemlidir. Bu teknoloji hemen her iş kolunda performansın artmasını, maliyetlerin düşmesini, işlerin hızlandırılmasını sağlayacağı gibi kitap, tez, makale gibi eserlerin oluşturulmasını kolaylaştıracağından sanatın, bilim ve teknolojinin gelişmesine de katkıda bulunacaktır.

KONUŞMA TANIMA

Konuşma Tanıma süreci mikrofondan alınan sesin sayılara dönüştürülmesiyle başlar. Bir ses algılama algoritması, sayısallaştırılan sinyali inceleyerek konuşmanın nerede başladığını ve bittiğini tespit eder. Algılanan konuşma parçası tanımada kullanılacak parametrelerin hesaplanması için çeşitli Sayısal Sinyal İşleme (Digital Signal Processing) algoritmalarından geçirilir. Söz konusu parametreler, Dil modeli ile desteklenmiş, bir tür yapay zeka algoritması olan tanıma algoritması tarafından konuşma tanıma için kullanılır.



Tanıma sürecinde kritik olan nokta tanınacak olan kelimenin önceden bilinmesinin zorunlu olmasıdır. Bilgisayar üzerindeki tanıma işlemleri gerçekte karşılaştırma işlemleri olup bu tür bilgisayar yazılımları önceden seçenek listesinde olmayan bir sonuca ulaşamazlar. Türkçe için zorluk bu noktadadır.

TÜRKÇE

Türkçe eklemeli yapısı nedeniyle çok üretken bir dildir. Sadece bir kök kelimeden yapım ve çekim ekleri kullanılarak milyonlarca yeni kelime türetmek mümkündür.

Bu çeşitlilik dilimize çok büyük bir ifâde gücü kazandırır. Yapım ve çekim ekleri almış pek çok kelimeyi bir başka dile tercüme etmek için cümle kurmanız gerekmektedir.

Aşağıdaki tabloda Türkçenin çok daha az sembolle çok daha fazla bilgi taşıdığı ayrıca bir kök kelimeden çok farklı anlamlarda yeni kelimeler üretebildiği açıkça görülmektedir.

"al_" Kökünden Türetilmiş Kelimeler	**İngilizce Karşılığı**
Alsam	I wish I take
Aldıysan	If you had taken
Alacaklı	Creditor
Almalıydım	I would have taken
Alıcı	Buyer
Alabilirsem	If I can take
Almalıysanız	If you should take
Alıcısızlıkla	With having no buyer (client)
Alıcılaşabilenlerdendi	He was one of those that was able to become buyer

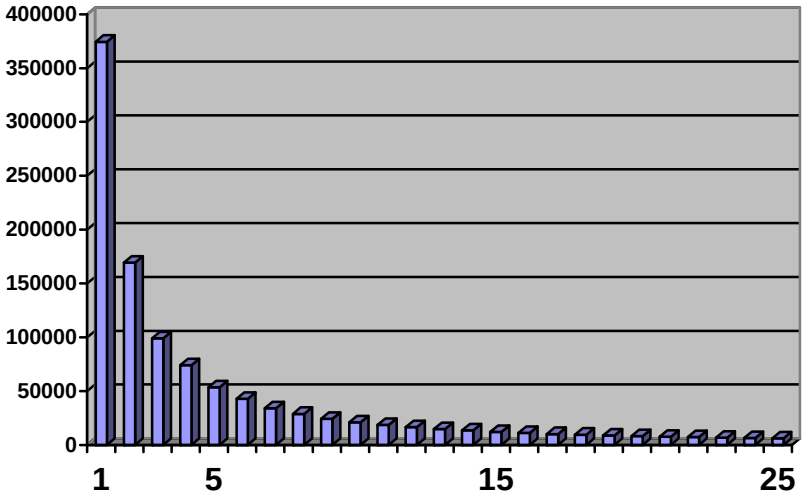
Dinleyici, kök kelimenin anlamını biliyorsa, Türkçe kurallara uygun türetilerek söylenen kelimeyi hayatında hiç duymamış dahi olsa anlar. Bu durum gerçekte çok sık karşılaşılan bir durumdur.

Örnek: Bugün için arama motoru olan Google'da "zürriyetsizmiş" kelimesini ararsanız yalnızca 1 defa bulunacaktır. Bir kelimenin internette oluşturulmuş tüm metinlerde yalnızca bir defa geçmesi bir istisna gibi görünebilir ancak aşağıda anlatılan çalışma bu durumun aslında çok sık rastlanılan bir gerçek olduğunu açıkça göstermektedir.

Türetilmiş kelimelerin görülme sıklıklarını incelemek için toplam 640 Milyon kelimeden (Yaklaşık 6gb) oluşan bir Türkçe metin verisi incelenmiştir. Birbirinden farklı kelime sayısının 1,410,000 adet olduğu görülmüştür.

Aşağıdaki grafiğe bakınız. Burada biraz önceki örnekte anlatılan durum açıkça görülmektedir. Birbirinden farklı 1,410,000 kelimenin 370,000 tanesi yalnızca 1 defa görülmüştür. İncelenen veri tabanında 2 defa tekrar eden kelime sayısı yaklaşık 170,000, 3

defa tekrar eden kelime sayısı yaklaşık 100bindir. Tekrar sayısı arttıkça kelime sayısı hızla azalmaktadır. 25 defa tekrar eden kelime sayısı 5bin iken 100 defa tekrar eden kelime sayısı sadece 885'dir.



Eğer türetilmiş kelimelerin hepsi günlük hayatta çok sık kullanılıyor olsaydı bunları metinlerden tespit etmek mümkün olabilirdi. Ancak görüldüğü gibi 640 milyon kelimeden oluşan bir veri tabanındaki birbirinden farklı 1,410,000 kelimenin yaklaşık %45'i 3 veya daha az sayıda görülmüştür.

Görüldüğü gibi metin toplayarak kullanılabilecek kelimeleri tespit etmek mümkün değildir. Bu nedenle konuşma tanıma projesinde; Türkçe ek ekleme kurallarına uygun olarak kelime türeten bir dil modeli oluşturulmuştur. Tanıma algoritması yeterince hızlı ve doğru olduğu için kullanılmayan türetilmiş kelimeler bir sorun teşkil etmemektedir.

Konuşma tanıma sisteminde kullanmak üzere Türkçe için geliştirdiğimiz dil modeli bir tane isim kökten 15 Milyon adet farklı yazılışta kelime türetebilmektedir. Sık kullandığımız kök kelime sayısının 25 bin civarında olduğunu düşünürseniz. Muhtemel türetilmiş kelime sayısının 200 Milyardan fazla olduğu ortaya çıkar. Geliştirilen dil modeli teorik olarak birbirinden farklı 300 milyar kelime sentezleyebilecek durumdadır.

GELİŞTİRİLEN KONUŞMA TANIMA TEKNOLOJİSİ

Konuşma tanıma sistemleri kelime kapasitelerine göre sınıflandırılırlar. Çünkü tanınan temel birim kelimedir. Kelime kapasitesi artırıldığında belirsizlik artacağı için doğru orantılı olarak hata yüzdesi de artar, ayrıca doğru orantılı olarak tanıma süresi ve işlemci yükü de artar.

Dünya üzerindeki en büyük sistemlerin kapasitesi 100,000 kelime civarındadır. Bu bildiride anlatılan sistem gerçek zamanlı olarak 300 Milyar kelime tanıma kapasitesine sahiptir. Söz konusu sistem 3 milyon kat daha yüksek bir iş zorluğu ile karşı karşıya olmasına rağmen tanıma doğruluğu ve hızı yönünden diğer sistemlerden geri kalmayan bir performansa sahiptir.

Bu bildiride anlatılan proje 2001 yılında başlamış olup 2004 yılında Dünya üzerindeki bilinen konuşma tanıma teknolojisinin seviyesi yakalanmıştır. 2004 yılında ulaşılan kelime kapasitesi 100 bindir. 2004'den sonraki 4 yıl tamamen araştırma ve yeni teknolojiler geliştirme ile geçmiştir. Bu süreç içinde sistemin çalışmasını sağlayacak yeni algoritmalar ve modeller icat edilmiştir.

Kaynaklar :

Muharrem Ergin, **Türk Dil Bilgisi** , İstanbul 1983.
Zeynep Korkmaz, **Türkiye Türkçesi Grameri** (Şekil Bilgisi), Ankara 2003.
Türkçe Sözlük, Türk Dil Kurumu Yayınları :549, Ankara 2005.
Yazım Kılavuzu, Türk Dil Kurumu Yayınları :859, Ankara 2005.
Tuncer Gülensoy, **Türkiye Türkçesindeki Türkçe Sözcüklerin Köken Bilgisi Sözlüğü,** Türk Dil Kurumu Yayınları :911, Ankara 2007, 2 Cilt
Tahsin Banguoğlu, **Türkçenin Grameri** (tıpkıbasım), Ankara 2007.

Geniş Dağarcıklı Türkçe Konuşma Tanıma İçin Yeni Bir Kod Çözme Algoritması *Onur Çilingir, Mübeccel Demirekler*

TREN- TURKISH SPEECH RECOGNITION PLATFORM
Hasan Palaz, Alper Kanak, Yücel Bicil, Mehmet Uğur Doğan, Tuba İslam

Türkçe Gazete Haberleri Dikte Sistemi
Ebru Arısoy, Levent M. Arslan