Pamukkale Univ Muh Bilim Derg, 24(2), 180-184, 2018

Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi

Pamukkale University Journal of Engineering Sciences

Mobil araçlarda Türkçe konuşma tanıma için yeni bir veri tabanı ve bu veri tabanı ile elde edilen ilk konuşma tanıma sonuçları

A new database for Turkish speech recognition on mobile devices and initial speech recognition results using the database

Osman BÜYÜK[1*]

[1]Elektronik ve Haberleşme Mühendisliği Bölümü, Mühendislik Fakültesi, Kocaeli Üniversitesi, Kocaeli, Türkiye.
osman.buyuk80@gmail.com

Geliş Tarihi/Received: 06.07.2016, Kabul Tarihi/Accepted: 24.10.2016
doi: 10.5505/pajes.2016.43765
* Yazışılan yazar/Corresponding author
Araştırma Makalesi/Research Article

Öz

Konuşma tanıma teknolojisi konuşmanın otomatik olarak metne dönüştürülmesini sağlamaktadır. Bu konuda yapılmış önceki çalışmalar, teknolojinin belli bir olgunluğa ulaşmasını ve pek çok farklı alanda kullanılmasını sağlamıştır. Son zamanlarda akıllı telefon, tablet gibi mobil uygulamaların kullanımında görülen hızlı artış konuşma tanıma teknolojisinin mobil platformlara uyarlanmasını önemli hale getirmiştir. Bu çalışmada mobil platformlar için yüksek başarım ile çalışan Türkçe bir konuşma tanıma sisteminin gerçekleştirilmesi hedeflenmiştir. Bu amaçla farklı akıllı telefonlardan alınmış kayıtlardan oluşan yeni bir ses veri tabanı oluşturulmuştur. Sistemin performansı üç farklı konuşma tanıma uygulaması kullanılarak ölçülmüştür. i) Televizyon kumanda uygulaması, ii) Sesli mesaj uygulaması, iii) Genel metin yazdırma uygulaması. Yaptığımız testlerde tanıma performansının televizyon kumanda uygulaması için %95'in üzerinde olduğu görülmüştür. Sesli mesaj ve genel metin yazdırma uygulamalarında yaklaşık %40 ve %60 başarım oranları elde edilmiştir.

Anahtar kelimeler: Ses işleme, Konuşma tanıma, Mobil araçlar

Abstract

The aim of speech recognition is to recognize human speech and convert it to written text. Past works in speech recognition technology leaded to significant improvements and this ensured the use of the technology in various practical applications. Recently, the demand for mobile applications has significantly increased when the smart phones and tablets have been introduced to the market. As a result, the adaptation of speech recognition to mobile devices has been an important issue since the technology has many applications in these devices. In this study, we aim to develop a Turkish speech recognition system for mobile devices. For this purpose, we collected a new database that includes recordings from various different speakers and smart phones. The performance of this system is tested using three speech recognition applications; i) Television control ii) Short message iii) General text dictation. In the experiments, we achieved 95% recognition performance in the grammar based television control application. The performance in short message and general text dictation applications are approximately %40 and %60, respectively.

Keywords: Speech processing, Speech recognition, Mobile devices

1 Giriş

Konuşma tanıma konuşmacının söylediği cümleleri en yüksek doğrulukla metne dönüştürme işlemi olarak tanımlanabilir. Konuşma tanıma alanındaki çalışmalar özellikle 1950'li yılların sonundan itibaren hız kazanmıştır. Modern konuşma tanıma sistemlerinde öznitelik çıkarımı, akustik model, dil modeli, tanıma sözlüğü ve tanıma algoritması modülleri bulunur.

Bir konuşma tanıma sistemindeki ilk adım sesin özniteliklerinin çıkartılmasıdır. Öznitelik çıkarımı sırasında ses işaretinin yavaş değişen özelliklerini yakalamak için, işaret genellikle ufak pencerelere bölünür. Bu pencereler birbiriyle kesişecek şekilde kaydırılır. Literatürde, öznitelik çıkarımı için çok sayıda yöntem önerilmiştir. Bunlar arasında mel-frekans kepstral katsayıları (mel-frequency cepstral coefficients-MFCCs) [1], lineer tahmin kepstral katsayıları (linear prediction cepstral coefficients-LPCCs) [2], algısal lineer tahmin katsayıları (perceptual linear predictive coefficients - PLPs) [3] en çok kullanılan yöntemlerdir.

Saklı Markov modelleri (hidden Markov model-HMM) stokastik süreçleri modellememizi sağlayan bir yöntemdir. Modern konuşma tanıma sistemlerinde HMM'ler akustik modelleme için en sık kullanılan yöntemdir [4]. HMM'ler konuşma tanımaya ek olarak el yazısı tanıma, jest tanıma, konuşmacı tanıma gibi diğer örüntü tanıma uygulamalarında da sıklıkla kullanılmaktadır. Konuşma tanıma uygulamalarında genelde hedef dildeki her ses birimi için ayrı bir HMM eğitilir. HMM eğitimi sırasında güvenilir modellerin elde edilmesi için fonetik olarak dengeli, farklı cinsiyet, yaş ve aksan özelliklerini içeren geniş bir ses veri tabanının kullanılması kritik öneme sahiptir. Bu veri tabanının uygulamanın kullanılacağı ortama uygun kayıtlardan oluşması sistemin performansını doğrudan etkilemektedir.

Dil modellemede ise uygulamanın tipine göre, basit kural tabanlı bir gramer ya da istatiksel N-gram dil modelleri tercih edilmektedir. Gramer tabanlı uygulamaların kompleksliği nispeten düşüktür. Bu durum hem tanıma performansı hem de tanıma hızını olumlu etkilemektedir. Diğer taraftan, bu uygulamalarda kullanıcının gramer tarafından tanımlanmış

cümleleri söylemesi beklenmektedir. Geniş dağarcıklı sürekli ses tanıma uygulamalarında (large vocabulary continous speech recognition-LVCSR) ise kullanılacak cümleler için böyle bir kısıtlama yoktur. LVCSR uygulamalarında, N-gram istatistiksel dil modelleri en fazla kullanılan yöntemlerdendir. Bu dil modellerinde sözlükte bulunan birimlerin birbirini izleme olasılıkları büyük bir metin verisinden tahmin edilmektedir. N-gram dil modellerinde her kelimenin (N-1) öncesine bakılarak bu olasılıklar hesaplanmaktadır.

Tanıma sözlüğünde, dil modelindeki her kelimenin akustik modelde bulunan ses birimleri ile nasıl oluşturulacağı bilgisi bulunmaktadır. Türkçe'de yazılış ve söyleyiş arasında yaklaşık birebir örtüşme olması tanıma sözlüğü oluşturulmasını oldukça kolaylaştırmaktadır. Tanıma algoritması ise tüm bu modülleri kullanarak ses işaretine en uygun kelime dizisini bulmaktadır.

Son yıllarda akıllı telefon gibi mobil cihazların kullanımı çok hızlı bir şekilde artmıştır. Örneğin bir mobil cihaz kullanıcısı günde yaklaşık olarak iki sa. telefonuna bakmaktadır. Mobil cihaz kullanımının artmasıyla birlikte herkes bilgiye daha hızlı, güvenilir ve rahat bir şekilde ulaşmak istemektedir. Bu noktada konuşma tanıma teknolojisi de kullanıcıların mobil cihazlarla daha doğal iletişime geçmelerini sağlayacak önemli bir uygulamadır. Mobil cihazlarda geliştirebilecek konuşma tanıma uygulamaları arasında;

i) Televizyon kumanda uygulaması,

ii) Sesli mesaj uygulaması,

iii) Genel metin yazdırma uygulaması sayılabilir. Herhangi bir kumandaya ihtiyaç duymadan televizyonu uzaktan ses ile kontrol edebilmek, araç kullanırken ya da elimiz doluyken hiç bir tuşa dokunmadan mesaj gönderebilmek veya telefonumuzla istediğimiz bir metni yazdırabilmek birçok mobil cihaz kullanıcısının günlük işlerini kolaylaştıracaktır.

Bu çalışmada bu üç uygulama mobil cihazlar için gerçekleştirilmiştir. Uygulamalarda en yüksek başarımı elde etmek için farklı mobil cihazlardan alınmış yeni bir ses veri tabanı oluşturulmuştur. Veri tabanında toplam 300'er cümleden oluşan iki ayrı cümle seti bulunmaktadır. Bu cümleler Türkçe'deki farklı ses birimlerini olabildiğince kapsayacak şekilde seçilmiştir. Cümleleri farklı aksan ve ses özelliklerine sahip yaklaşık 500 farklı kişi tekrar etmiştir. Kayıtlar sırasında farklı akıllı telefonlar kullanılmıştır. Böylece 10 sa.'lik mobil konuşma tanıma uygulamalarında kullanılabilecek yeni bir ses veri tabanı oluşturulmuştur. Bu veri tabanı ile HMM eğitimi gerçekleştirilmiştir. Bu HMM'ler TV kumanda, sesli mesaj ve genel metin yazdırma uygulamaları için test edilmiştir. TV kumanda uygulaması için gramer tabanlı, diğer iki uygulama için ikili (bi-gram) dil modeli kullanılmıştır. Testlerde TV kumanda uygulaması için %95'in üzerinde tanıma performansı elde edilmiştir. Sesli mesaj ve genel metin yazdırma uygulamaları için başarım oranları sırasıyla yaklaşık %40 ve %60 olarak gerçekleşmiştir.

Çalışmanın kalan kısmı şu şekilde düzenlenmiştir: İkinci bölümde, bir konuşma tanıma sisteminin modülleri ayrıntılarıyla anlatılacaktır. Üçüncü bölüm ses ve metin veri tabanlarına ayrılmıştır. Dördüncü bölümde yapılan deney sonuçları paylaşılacaktır. Son bölümde gelecekte yapılması düşünülen çalışmalar belirtilecektir.

2 Konuşma tanıma

Bir konuşma tanıma sistemi öznitelik çıkarımı, akustik model, dil modeli ve tanıma algoritması modüllerinden oluşmaktadır. Bu modüllerin ayrıntıları izleyen alt bölümlerde anlatılacaktır.

2.1 Öznitelik çıkarımı

Bir konuşma tanıma sisteminin ilk aşamasında ses işaretinden konuşma tanıma için en ayırt edici öznitelikleri çıkarılır. En sık kullanılan öznitelik çıkarma yöntemlerinden birisi mel-frekans kepstral katsayıları (MFCC) metodudur. MFCC'ler hesaplanırken, pencerelenmiş ses işaretinin ayrık zamanlı Fourier dönüşümü (discrete-time Fourier transform-DFT) alınır ve DFT katsayıları frekans bantlarına ayırılır. Frekans bantlarındaki DFT değerleri genelde üçgen bir pencere ile çarpılarak, sonuçlar her frekans bandı için ayrı toplanmaktadır. Son aşamada, her banttaki büyüklük değerlerinin ayrık zamanlı kosinüs dönüşümü (discrete cosine transform-DCT) alınmaktadır. MFCC yönteminde, frekans bantları normal frekans skalasında değil, mel frekans skalasında doğrusal olarak dağıtılmıştır. Mel frekans skalası ise aşağıdaki şekilde tanımlanmıştır [5];

$$Mel(f) = 2595 * \log(1 + \frac{f}{700}) \quad (1)$$

Konuşma tanıma uygulamalarında genellikle, öznitelik vektörü olarak DCT sonrası elde edilen ilk 12 katsayı ile birlikte enerji değeri de kullanılmaktadır. Böylece, 13 boyutlu öznitelik vektörü oluşturulmaktadır. Yapılan çalışmalarda MFCC vektörlerine birinci ve ikinci derece regresyon katsayılarının eklenmesinin tanıma performansını olumlu etkilediği gözlenmiştir. Sonuç olarak, tipik bir konuşma tanıma uygulamasında birinci ve ikinci derece regresyon katsayıları ile birlikte 39 boyutlu öznitelik vektörü kullanılmaktadır.

2.2 Akustik model

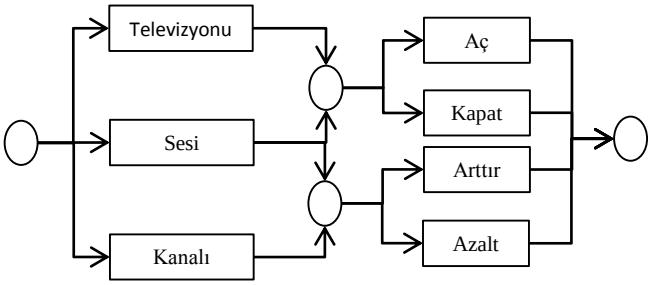
Modern konuşma tanıma sistemlerinde hedef dildeki her ses birimi bir HMM ile modellenmektedir. HMM modellemesi sırasında artikülasyon bilgisinden faydalanarak tanıma başarımını arttırmak için, konteks bağımsız (context independent) modeller yerine, konteks bağımlı (context dependent) modeller tercih edilmektedir. Üçlü (tri-phone) konteks bağımlı modelller her bir ses biriminin sağ ve sol konteksi göz önünde bulundurularak eğitilmektedir. Eğitim sırasında, birbirine benzeyen üçlü model durumları sınıflanarak, birbirine bağlanmaktadır. Bu işlem, hem eğitilecek üçlü model sayısını hem de bu modelleri eğitmek için gerekli olan eğitim verisi miktarını azaltmaktadır. Bu yöntemde her HMM durumu bir Gauss karışım modeli (Gaussian mixture model-GMM) ile modellenir. Her ses birimi için 3 durum, her durum için 12 karışım sık tercih edilen parametrelerdir.

2.3 Dil modeli

Konuşma tanıma uygulamaları, uygulamanın tipine göre gramer tabanlı ve geniş dağarcıklı olarak iki kategoriye ayrılabilir. Şekil 1'de televizyondaki temel fonksiyonları ses ile kumanda etmek için kullanılabilecek basit bir sonlu durum gramer (finite state grammar-FSG) yapısı gösterilmiştir;

Şekil 1'de görüldüğü gibi, gramer tabanlı uygulamaların kompleksliği nispeten düşüktür. Bu durum hem tanıma performansı hem de tanıma hızını olumlu etkilemektedir. Diğer taraftan, bu uygulamalarda kullanıcının gramer

tarafından tanımlanmış cümleleri söylemesi beklenmektedir. Örneğin yukarıdaki örnekte, kullanıcının "sesi arttır" yerine "sesi yükselt" komutunu kullanması durumunda, bu komutun sistem tarafından tanınma olasılığı yoktur. Geniş dağarcıklı sürekli ses tanıma uygulamalarında (LVCSR) ise kullanılacak cümleler için böyle bir kısıtlama yoktur. LVCSR uygulamalarında, N-gram istatistiksel dil modelleri en fazla kullanılan yöntemlerdendir. Bu dil modellerinde sözlükte bulunan birimlerin (çoğunlukla kelimeler) birbirini izleme olasılıkları büyük bir metin verisinden tahmin edilmektedir. N-gram dil modellerinde her kelimenin (N-1) öncesine bakılarak bu olasılıklar hesaplanmaktadır. Konuşma tanıma sistemlerinde hız/performans dengesi açısından genelde ikili (bi-gram) ya da üçlü (tri-gram) dil modelleri tercih edilmektedir.



Şekil 1: Ses ile TV kumandası için sonlu durumlu gramer örneği.

2.4 Konuşma tanıma algoritması

Tanıma algoritmasında, akustik özniteliklere en uygun kelime dizisi bulunmaya çalışılır. Bu problem aşağıdaki şekilde ifade edilebilir (Rabiner ve Huang, 1993);

$$\widehat{W} = w_1 w_2 \dots w_n = \underset{W}{\operatorname{argmax}}\, P(W|A) \tag{2}$$

Denklemde, A akustik vektörleri ve $\widehat{W}$ ise akustik vektörlere en uygun kelime dizisini ifade etmektedir. Bayes kuralı kullanılarak, Denklem 2 aşağıdaki şekilde yazılabilir;

$$\widehat{W} = \underset{W}{\operatorname{argmax}} \frac{P(A|W)P(W)}{P(A)} = \underset{W}{\operatorname{argmax}}\, P(A|W)P(W) \tag{3}$$

Denklem 3'te $P(A|W)$ olasılıkları akustik modelden, $P(W)$ olasılıkları ise dil modelinden elde edilmektedir. Dil modeli ve akustik model olasılıkları kullanılarak ses verisine en uygun kelime dizisi Viterbi algoritması kullanılarak bulunmaktadır [6].

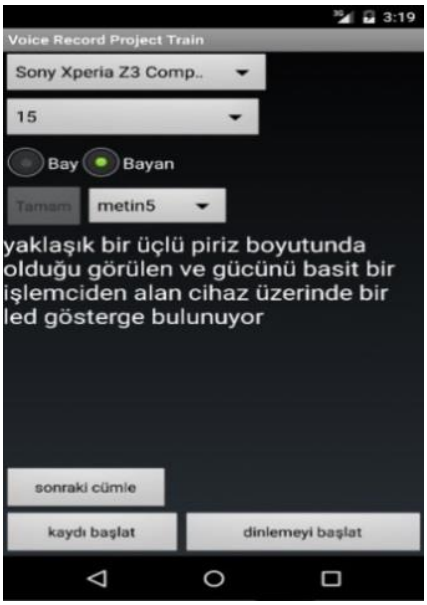
Konuşma tanıma performansını olumsuz etkileyen en önemli faktörlerden birisi yüksek geri-plan gürültüsüdür. Gürültülü ortamlardaki tanıma performansını arttırmak için literatürde çok sayıda yöntem önerilmiştir. Bu yöntemlerden bazıları öznitelik çıkarma modülünden önce bir ön işlem olarak uygulanmaktadır. Bu çalışmada mobil cihazlardan alınan kayıtlarda geri-plan gürültüsünün bazı durumlarda yüksek olması sağlanmıştır. Böylece yüksek geri plan gürültüsünde performans analizi de yapılmıştır. Yüksek geri plan gürültüsündeki tanıma sonuçları 4. bölümde verilmiştir.

3 Ses ve metin veri tabanları

Konuşma tanıma sistemlerinde akustik model eğitimi için geniş ses veri tabanlarına ihtiyaç duyulmaktadır. Bu verilerin uygulamanın pratikte kullanılacağı ortama uygun toplanması tanıma performansını olumlu etkilemektedir. Bilgisayar veya telefon üzerinden Türkçe konuşma tanıma uygulamalarında kullanılabilecek büyük ses veri tabanları bulunmasına rağmen, bildiğimiz kadarıyla mobil platformlara özel bir veri bulunmamaktadır. Bu nedenle, bu çalışmanın ilk aşamasında mobil konuşma tanıma için geniş ses veri tabanının oluşturulması üzerinde çalışılmıştır. Ayrıca geniş dağarcıklı uygulamalarda istatiksel dil modeli eğitimi için metin verisine ihtiyaç duyulmaktadır. İzleyen alt bölümlerde bu ses ve metin verilerinin özellikleri anlatılacaktır.

3.1 Mobil konuşma tanıma veri tabanı

Çalışmanın başlangıcında çok sayıda konuşmacının ses kayıtlarını doğru ve hızlı bir şekilde alabilmek için Android platformunda bir ara yüz geliştirilmiştir. Bu ara yüz Şekil 2'de görülebilir;



Şekil 2: Konuşma tanıma veri tabanı ses kaydı toplama ara yüzü.

Konuşma tanıma uygulaması için topladığımız veri tabanı gazetelerden, dergilerden ve belgesellerden alınan toplam 600 cümleden oluşmaktadır. Bu cümlelerden seçtiğimiz bazı örnekler aşağıda verilmiştir;

Örnek eğitim cümleleri;

"son yıllarda özellikle sosyal medyada kitap yüzü gördükçe içimiz bir mest oluyor"

"kitap kurtları için kılavuz niteliğinde bir listeye hazır mısınız"

"stresli misin uzmanlar stresi alt etmenin on şaşırtıcı yolunu açıkladı"

Her setteki 300 cümle onar cümlelik daha küçük parçalara ayrılmıştır. Veri tabanına ses kaydı vermek isteyen konuşmacılardan kendileri için belirlenen on cümleyi Şekil 2'de ekran görüntüsü verilen ara yüzü kullanarak tekrar etmeleri istenmiştir. Ara yüzde görüldüğü gibi kayıt veren konuşmacıyı yönlendirmek için okunacak cümle ayrıca telefon ekranında gösterilmektedir. Konuşmacılar çoğunlukla sadece bir adet on cümlelik parçayı tekrar etmiştir. Bazı konuşmacılar birden fazla on cümlelik parça okuyarak çalışmaya katkıda bulunmuşlardır. Farklı söyleyiş biçimleri ve ses özelliklerini veri tabanında kapsayabilmek için kayıtlar olabildiğince farklı

konuşmacıdan alınmıştır. Veri tabanı için kayıt veren yaklaşık 500 farklı konuşmacı bulunmaktadır. Bu konuşmacıların cinsiyet, yaş gibi özelliklerinin dengeli dağılmasına dikkat edilmiştir. Ayrıca, gelecekte yapılacak çalışmalara temel teşkil etmesi için konuşmacıların cinsiyet ve yaş bilgileri de alınmıştır. Ses kayıtları farklı akıllı telefonlar kullanılarak alınmıştır.

Şu ana kadar topladığımız veri tabanında toplam yaklaşık 10 sa.'lik ses verisi bulunmaktadır. Gelecekte bu veri miktarını arttırmayı planlıyoruz. Ayrıca topladığımız veri tabanını akademik araştırma amacıyla kullanıma açmayı planlıyoruz. Veri tabanındaki tüm kayıtlar 16 kHz, 16 bit, tek kanal (mono), darbe kod modülasyonu (pulse code modulation–PCM) formatında alınmıştır. PCM sıkıştırılmamış bir format olması nedeniyle tercih edilmiştir.

3.2 Metin veri tabanları

LVCSR uygulamalarında, dil modellemesinde istatistiksel N-gram dil modelleri kullanılmaktadır. N-gram dil modellerinde kelimelerin birbirini izleme olasılıklarını güvenilir bir şekilde hesaplamak için büyük metin verilerine ihtiyaç duyulmaktadır. Bu metin verileri için internet önemli bir kaynaktır. Bu çalışmada internetten metin verilerini otomatik olarak indirip, dil modeli eğitmek için JAVA dilinde basit bir ara yüz geliştirilmiştir.

Bu ara yüzde girdi olarak alınan web sayfalarının içerikleri bir metin dosyasına indirilmektedir. İlk indirilen metin verileri genelde önemli ölçüde gürültü içermektedir. İkinci aşamada bu metin verileri temizlenmekte ve istenmeyen içeriklerden arındırılmaktadır. Bu aşamada metin verileri cümlelerine de ayrılmaktadır. Temizleme aşamasından sonra bu metinlerle genel metin yazdırma uygulaması için istatiksel dil modeli eğitimi yapılmaktadır. Bu çalışmada genel metin yazdırma dil modeli eğitimi için yaklaşık 4 milyon cümle kullanılmıştır.

Sesli mesaj yazdırma uygulaması için farklı kişilerden örnek mesaj cümleleri toplanmıştır. Ayrıca, internetten bazı mesaj şablonları bulunarak eğitim verisine eklenmiştir. Böylece yaklaşık 1000 cümleden oluşan bir kısa mesaj metin verisi oluşturulmuştur. İstatistiksel dil modeli eğitimi bu metin verisi kullanılarak yapılmıştır.

4 Deneyler

Çalışmanın ilk aşamasında TV kumandası için gramer tabanlı bir uygulama geliştirilmiştir. Gramer tabanlı uygulama seçilmesinde, bu uygulamalardaki yüksek başarım oranı etkili olmuştur. Çalışmanın ikinci aşamasında iki farklı LVCSR uygulaması geliştirilmiştir. Bu uygulamalardan ilki sesli mesaj yazdırma, ikincisi genel metin yazdırma uygulamasıdır. Bu uygulamalarda dil modellemesi için istatistiksel ikili dil (bi-gram) modeli kullanılmıştır. Bu dil modelleri Bölüm 3.2'de anlatılan metin verileri kullanılarak eğitilmiştir.

Farklı konuşma tanıma uygulamalarında kullanılmak üzere geliştirilmiş açık kaynak kodlu kütüphaneler bulunmaktadır [5],[7],[8]. Bu kütüphanelerden birçok araştırma çalışmasında faydalanılmıştır. Ayrıca bu kütüphaneler kullanılarak farklı ürünler geliştirilmiştir [9]-[11]. Çalışmamızda konuşma tanıma uygulaması için açık kaynak kodlu Sphinx Kütüphanesi kullanılmıştır [7]. Sphinx kütüphanesinin seçilmesinde mobil platformlara uyarlanabilmesi etkili olmuştur. Sphinx kütüphanesi ile akustik model eğitimi ve testleri gerçekleştirilmiştir. LVCSR uygulamalarında istatistiksel dil modeli eğitimi için bir başka açık kaynak kodlu yazılım olan CMU SLM Kütüphanesi kullanılmıştır [12]. CMU SLM, Sphinx ile entegre çalışabilmesi nedeniyle seçilmiştir.

TV kumanda uygulaması için toplam 50 komuttan oluşan bir gramer oluşturulmuştur. Gramerdeki tüm komutlar birbiriyle eşit olasılığa sahiptir. Bu uygulamadaki başarımı ölçmek için eğitim verisinden farklı bir test verisi toplanmıştır. Test verisi için elli komut cümlesi onarlık gruplara ayrılmıştır. Her test konuşmacısı kendisi için belirlenen on cümlelik komut listesini okumuştur. Test verisinde yaklaşık 20 farklı kişi vardır. Bu test verisi ile elde ettiğimiz tanıma sonuçları Tablo 1'de verilmiştir. Tabloda görüldüğü gibi basit bir FSG uygulamasında tanıma başarımı oldukça yüksektir. Bu uygulamada kelime doğruluğu oranı yaklaşık %98 olarak gerçekleşmiştir.

Tablo 1: TV kumanda uygulaması konuşma tanıma başarım oranları.

	%Tanıma Oranı
Cümle Tanıma Doğruluğu	96.00
Kelime Tanıma Doğruluğu	98.08
Kelime Tanıma Kesinliği	98.08

Konuşma tanıma sisteminin genel metin yazma uygulamasındaki başarımını ölçmek için gazetelerin internet sitelerinden 100 farklı test cümlesi belirlenmiştir. Bu cümleler onarlık gruplara ayrılmıştır. Yaklaşık 20 farklı kişi belirlenen cümleleri okumuştur. Bu test verisi ile elde ettiğimiz tanıma oranları Tablo 2'de verilmiştir. Tabloda görüldüğü gibi genel metin yazdırma uygulamasında yaklaşık %60 kelime tanıma oranı elde edilmektedir. Bu oran daha önce Türkçe için yapılmış benzer çalışmalardaki tanıma oranları ile uyumludur [10],[11],[13].

Tablo 2: Genel metin yazdırma uygulaması konuşma tanıma başarım oranları.

	%Tanıma Oranı
Cümle Tanıma Doğruluğu	9.70
Kelime Tanıma Doğruluğu	61.30
Kelime Tanıma Kesinliği	50.75

Sesli mesaj uygulamasındaki başarımı ölçmek için 50 örnek kısa mesaj cümlesi toplanmıştır. Bu cümleler 4 farklı kişi tarafından gürültülü ve gürültüsüz ortamda okunmuştur. Gürültüsüz kayıtlar sessiz ofis ortamında, gürültülü kayıtlar kalabalık sınıf ortamında alınmıştır. Bu test verisi ile elde ettiğimiz tanıma oranları Tablo 3'te verilmiştir.

Tablo 3: Sesli mesaj uygulaması konuşma tanıma başarım oranları.

	Gürültüsüz Ortam	Gürültülü Ortam
Cümle Tanıma Doğruluğu	5.50	2.00
Kelime Tanıma Doğruluğu	40.69	29.44
Kelime Tanıma Kesinliği	21.25	11.81

Tablo 3'te görüldüğü gibi sesli mesaj uygulamasındaki tanıma başarımı, daha kısıtlı bir alan olmasına rağmen, genel metin yazma uygulamasındaki başarımdan düşüktür. Bunun en önemli sebebi sesli mesaj dil modeli eğitiminde kullanılan metin miktarının az olmasıdır. Gelecekte sesli mesaj metin verisinin genişletilmesi planlanmaktadır. İkinci sebep, kısa mesajların kişiye özgü kelimeler ve kısaltmalar içermesidir. Bu durum bu uygulamalardaki sözlük dışı kelime oranının artmasına neden olmaktadır. Sesli mesaj uygulamasının mobil platformlarda yüksek performans ile çalışabilmesi için, dil modelinin kullanıcının yazdığı kısa metin mesajlarıyla adapte

edilmesi gerekmektedir. Son olarak, yüksek geri plan gürültüsü beklendiği gibi tanıma performansını olumsuz etkilemektedir. Yüksek geri plan gürültüsü ile kelime doğruluk oranı %40'lardan %30'a düşmektedir.

5 Sonuçlar ve gelecek çalışmalar

Bu çalışmada mobil platformlar için bir konuşma tanıma uygulaması gerçekleştirilmiştir. Bu amaçla farklı akıllı telefonlardan alınmış kayıtlardan oluşan yeni bir veri tabanı oluşturulmuştur. Bildiğimiz kadarıyla Türkçe mobil konuşma tanıma uygulamaları için daha önce toplanmış böyle bir veri tabanı bulunmamaktadır. Bu nedenle topladığımız veri tabanının bu alanda yapılacak çalışmalara önemli bir katkı olacağını düşünüyoruz. Gelecekte bu veri tabanını akademik araştırma amacıyla kullanıma açmayı planlıyoruz.

Topladığımız mobil konuşma tanıma veri tabanı gelecekte yapacağımız çalışmalar için önemli bir temel teşkil edecektir. Bu nedenle veri tabanını gelecekte genişletmeyi planlıyoruz. Veri tabanında konuşmacıların kayıtlarına ek olarak cinsiyet ve yaş bilgileri de bulunmaktadır. Bu bilgilerin konuşmadan yaş ve cinsiyet tespiti gibi uygulamalarda kullanılabileceğini düşünüyoruz.

Bu çalışmada konuşma tanıma sisteminin performansı üç farklı mobil uygulamada test edilmiştir;

i. Gramer tabanlı televizyon kumanda uygulaması,
ii. Sesli mesaj uygulaması,
iii. Genel metin yazdırma uygulaması. Testlerde televizyon kumanda uygulaması gibi sınırlı sayıda komut içeren bir uygulamada %95'in üzerinde başarım elde edilirken, geniş dağarcıklı uygulamalarda tanıma performansı yaklaşık %50 olmuştur.

Konuşma tanıma uygulamalarının gömülü olarak mobil platformlara uyarlanması sırasında farklı zorluklarla karşılaşılmaktadır. Bunlardan en önemli ikisi mobil araçlardaki kısıtlı kaynak miktarı ve mobil kayıtlardaki yüksek geri plan gürültüsüdür. Tüm mobil araçlarda yüksek başarım ile çalışacak konuşma tanıma uygulamaları gerçekleştirmek için bu sorunların üzerinde durulması gerekmektedir.

Çalışmamızda yüksek geri plan gürültüsününün konuşma tanıma performansı üzerindeki etkisi incelenmiştir. Yaptığımız testlerde tanıma başarımının yaklaşık %40'tan yüksek geri plan gürültüsü ile %30'lara düştüğü gözlenmiştir. Geri plan gürültüsünün yüksek olduğu ortamlarda ses kalitesini iyileştirmek için yaptığımız çalışmalar devam etmektedir. Ayrıca, gelecekte veri tabanını yüksek geri plan gürültülü kayıtlarla genişletmeyi planlıyoruz. Farklı ortamlardan alınmış kayıtların tanıma performansını arttıracağını düşünüyoruz.

Mobil araçlardaki kısıtlı kaynaklar yüksek kaynak gerektiren LVCSR gibi uygulamaların gerçekleştirilmesi önünde önemli bir engel teşkil etmektedir. LVCSR uygulamalarında kullanılan kaynak miktarını ve uygulamanın hızını temel olarak tanıma sözlüğündeki kelime sayısı belirlemektedir. Türkçe'nin eklemeli yapısı LVCSR uygulamalarında önemli bir zorluğa sebep olmaktadır. Bu uygulamalarda dildeki kelimelerin büyük bir kısmının tanıma sözlüğü tarafından kapsanacağı varsayılmaktadır. Sınırlı sayıda kelime içeren bir sözlüğün Türkçe'deki kelimelerin büyük bir kısmını kapsaması eklemeli yapı nedeniyle mümkün değildir. Bu durum, Türkçe LVCSR uygulamalarında sözlük dışı kelime oranının (out-of-vocabulary rate-OOV) yüksek olmasına sebep olmaktadır. Sözlükte olmayan kelimelerin sistem tarafından tanınma olasılığı bulunmadığından, yüksek sözlük dışı kelime oranı tanıma performansını olumsuz etkilemektedir. Kapsama oranını arttırmak için kelime sayısının arttırılması ise mobil uygulamalardaki kaynak kısıtı nedeniyle mümkün değildir. Literatürde, bu problemi gidermek için tanıma birimi olarak kelime altı (kök-ek ya da hece) birimlerin kullanılması önerilmiştir. Gelecekte geliştirdiğimiz mobil LVCSR uygulamalarında kelime altı birimlerin kullanılması da değerlendirilecektir.

6 Teşekkür

Bu çalışma TÜBİTAK 3001 programı çerçevesinde 114E742 No.lu proje kapsamında desteklenmiştir.

7 Kaynaklar

[1] Davis SB, Mermelstein P. "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences". *IEEE Transactions on Acoustics, Speech and Signal Processing*, 28(4), 357-366, 1980.

[2] Makhoul J. "Linear prediction: A tutorial review". *Proceeding of the IEEE*, 63(4), 561-580, 1975.

[3] Hermansky H, Morgan N, Bayya A, Kohn P. "RASTA-PLP speech analysis technique". *IEEE International Conference on Acoustics, Speech and Signal Processing 1992*, San Francisco, California, USA, 23-26 March 1992.

[4] Rabiner LR, Huang BW. *Fundamentals of Speech Recognition*. Englewood Cliffs, New Jersey, USA, Prentice Hall Inc, 1993.

[5] Young S, Evermann G, Gales M, Hain T, Kershaw D, Liu X, Moore G, Odell J, Ollason D, Povey D, Valtchev V, Woodland P. *The HTK Book (for HTK Version 3.4)*. 3th ed. Cambridge, UK, Cambridge University Engineering Department, 2006.

[6] Viterbi, AJ. "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm". *IEEE Transactions on Information Theory*, 13(2), 260-269, 1967.

[7] Carnegie Mellon University Speech Processing Group. "Carnegie Mellon University Sphinx, Open Source Toolkit for Speech Recognition". http://cmusphinx.sourceforge.net (19.04.2014).

[8] Kaldi Project. "Kaldi: A Toolkit for Speech Recognition". http://kaldi-asr.org/ (05.01.2016).

[9] Tan ZH, Lindberg B. "Speech recognition on mobile devices". *Lecture Notes in Computer Science*, 5960, 221-237, 2010.

[10] Arisoy E. Turkish Dictation System for Radiology and Brodcast News Applications. Msc. Thesis, Bogazici University, Turkey, 2004.

[11] Buyuk O. Sub-word Language Modelling for Turkish Speech Recognition. Msc. Thesis, Sabanci University, Turkey, 2005.

[12] Carnegie Mellon University Speech Processing Group, CMU. "The CMU Statistical Language Modeling (SLM) Toolkit". http://www.speech.cs.cmu.edu/SLM_info.html (19.10.2015).

[13] Buyuk O, Haznedaroglu A, Arslan LM. "Turkish speech recognition software with adaptable language model". *15th Signal Processing and Communication Applications Conference*, Eskisehir, Turkey, 11-13 June 2007.