DİL MODELİ UYARLANABİLİR TÜRKÇE SES TANIMA YAZILIMI

TURKISH SPEECH RECOGNITION SOFTWARE WITH ADAPTABLE LANGUAGE MODEL

Osman Büyük, Ali Haznedaroğlu, Levent M. Arslan

Elektrik ve Elektronik Mühendisligi Bölümü, Bogaziçi Üniversitesi, 34342, Bebek, İstanbul
{osman.buyuk, ali.haznedaroglu}@sestek.com.tr,
arslanle@boun.edu.tr

Özetçe

Son yıllarda, Türkçe için ses tanıma konusunda yapılan çalışmalar hız kazanmıştır. Bu çalışmalar çerçevesinde, ses tanıma uygulamalarında kullanılabilecek ses ve metin verisinde önemli artış görüldüğü gibi, günlük hayatta kullanılabilecek kadar başarımı yüksek bir yazılım gerçekleştirmek için önerilen metodların sayısı da artmıştır. Türkçe gibi eklemeli dillerdeki geniş dağarcıklı ses tanıma uygulamalarında karşılaşılan kapsama problemini çözmek için sözcük-altı birimlerin kullanımı önerilirken daha kısıtlı alanlarda yapılmış ses tanıma çalışmaları da bulunmaktadır. Bu makalede, son zamanlarda ivme kazanan çalışmalardan yararlanılarak geliştirilen bireysel kullanıcıya yönelik bir ses tanıma yazılımının tanıtımı yapılacaktır. Bu yazılımın arayüzü ile ilgili bilgiler verildiği gibi iki farklı alandaki tanıma performansı da özetlenecektir. Yazılımının performansı sınırlı bir alan olan radyoloji ile geniş dağarcıklı bir test verisinde sınanmıştır. Türkçedeki kapsama problemi ise sözlük ve dil modelinin kullanıcı tarafından sık kullanılan kelime ya da cümle kalıplarına adapte edilmesi sayesinde aşılmaya çalışılmıştır. Sınamalar sonucunda radyolojide yaklaşık %90, geniş dağarcıklı testte ise yaklaşık %44 kelime tanıma performansı elde edilmiştir.

Abstract

Turkish speech recognition studies have been accelerated recently. With these efforts, not only available speech and text corpus which can be used in recognition experiments but also proposed new methods to improve accuracy has increased. Agglutinative nature of Turkish causes out of vocabulary (OOV) problem in Large Vocabulary Continuous Speech Recognition (LVCSR) tasks. In order to overcome OOV problem, usage of sub-word units has been proposed. In addition to LVCSR experiments, there have been some efforts to implement a speech recognizer in limited domains such as radiology. In this paper, we will present Turkish speech recognition software, which has been developed by utilizing recent studies. Both interface of software and recognition accuracies in two different test sets will be summarized. The performance of software has been evaluated using radiology and large vocabulary test sets. In order to solve OOV problem practically, we propose to adapt language models using frequent words or sentences. In recognition experiments, 90% and 44% word accuracies have been achieved in radiology and large vocabulary test sets respectively.

1. Giriş

Eklemeli bir dil olan Türkçe geniş dağarcıklı ses tanıma uygulamalarında (Large Vocabulary Continous Speech Recognition – LVCSR) problem teşkil etmektedir. Kod çözücüye verdiğimiz önceden belirlenmiş tanıma sözlüğü, geniş dağarcıklı uygulamalarda üretilebilecek kelimelerin çokluğu nedeniyle istenilen kapsama oranını sağlayamamaktadır. Kelimelerin tanıma birimi olarak kullanıldığı LVCSR uygulamalarında, sınama verisindeki kelimelerin %15-20'sinin tanıma sözlüğünün dışında kaldığı görülmüştür [1],[2]. Kapsam dışı kelimelerin kod çözücü tarafından tanınma olasılıkları olmadığından, bu kelimeler performansı olumsuz yönde etkilemekte ve Türkçe için kelimelerin tanıma birimi olarak kullanıldığı LVCSR uygulamalarının İngilizce gibi dillere oranla başarısız olmasına neden olmaktadır. Türkçenin yapısından kaynaklanan kapsama problemini çözmek için sözcük-altı dil birimlerinin (hece, ek veya kök) kullanımı önerilmiştir. Sözcük altı birimlerin kullanımı ile kapsama probleminin çözüldüğü görülürken, tanıma oranlarında da artış sağlanmıştır [1],[2],[3],[4],[5],[6],[7],[8].

Türkçe LVCSR uygulamalarında kapsama problemi nedeniyle istenen performans elde edilememesine rağmen, daha sınırlı gramarlerin kullanılabileceği alanlarda yüksek başarım oranları elde edilebilir. Örneğin radyoloji raporları gibi kısıtlı cümle kalıplarının ve sözcük dağarcığının kullanıldığı bir alanda daha önce yapılmış ses tanıma uygulamasında yaklaşık %84 kelime tanıma performansı elde edilmiştir [2]. Bu tanıma performansı ile çalışacak bir ses tanıma yazılımı işyerlerinde genellikle meşgul olan radyologların radyoloji raporlarını metne çevirmesinde büyük kolaylıklar sağlayabilir.

Bu çalışmada hem geniş dağarcıklı hem de radyoloji gibi daha sınırlı ses tanıma uygulamalarında kullanılabilecek bir dikte yazılımının tanıtımı yapılacaktır. Bu yazılımda Türkçe'de karşılaşılan kapsama problemini çözmek için kullanıcının önceden eğitilmiş dil modellerini istediği kelime ya da cümle kalıplarıyla adapte etmesi sağlanmıştır. Kullanıcı sıklıkla kullandığı kelimeleri ya da belli bir alanda toplanmış metinleri dil modeline ekleyerek, daha önceden eğitilmiş ikili dil modelini

istediği alana uyarlayabilmektedir. Böylece tanıma sözlüğü kullanıcının isteğine göre genişleyebilen bir ses tanıma yazılımı gerçekleştirilmiştir.

Bu yazılımının performansını ölçmek için radyoloji alanında toplanmış ve günlük bir gazetenin farklı haber kategorilerinde geçen cümlelerden oluşturulmuş iki farklı ses verisi kullanıldı. Radyoloji alanında toplanan veri yazılımın sınırlı bir alanda göstereceği performansı ortaya koyarken, günlük gazetenin sayfasında geçen cümlelerden oluşturulmuş veri geniş dağarcıklı ses tanıma uygulamasındaki kullanılabilirliğini ortaya koyacaktır.

Bu bildiri şu şekilde düzenlenmiştir. Bölüm 2'de yazılımın gerçekleştirilmesinde kullanılan ses ve metin verilerinin özellikleri anlatılacaktır. Bölüm 3'te yazılımın arayüzü ile ilgili detaylar verilirken, Bölüm 4'te tanıma deneylerinden elde edilen sonuçlar aktarılacaktır. Bildirinin son bölümünde çalışmanın sonuçlarını ve geleceğe yönelik planları bulabilirsiniz.

2. Ses ve Metin Verisi Özellikleri

Geliştirilen yazılımda akustik modelleme için Saklı Markov Modelleri (Hidden Markov Models - HMM) kullanıldı. Bu modellerin eğitimi Sabancı Üniversitesi [6], Boğaziçi Üniversitesi [2] ve ODTÜ'de [9] toplanmış ses verileri kullanılarak yapıldı. Bu veri toplama işlemi sonucunda, Türkçe akustik model eğitiminde kullanılabilecek yaklaşık 53 saatlik ses veritabanı oluşturuldu. Bu verilerin tamamı 16 kHz'de örneklenirken, nicemleme için 16 bit kullanıldı.

Eğitim için ayrılmış ses verileri kullanılarak Türkçe alfabedeki 29 ses birimi için 5 durumlu ve 12 karışımlı metne bağımlı üçlü (context-dependent triphone) HMM'ler eğitildi. Üçlü modellerin eğitiminde bazı durumlara denk gelen verinin azlığından kaynaklanabilecek model parametrelerinin güvenilir elde edilememesi problemi Türkçe seslerin akustik benzerliği dikkate alınarak hazırlanan karar ağacı gruplandırması (decision tree clustering) ile aşılmaya çalışıldı [1]. Eğitilen ses birimleri ile Türkçe harfler arasında birebir eşleşme olduğu kabul edilirken, "ğ" harfi için de ayrı bir akustik model eğitildi. Akustik modellemede 29 birimin kullanılması yazılımın geçekleştirilmesinde kolaylık sağladığı için tercih edildi. Ayrıca bu seçimin tanıma performansını çok etkilemeyeceği düşünüldü. Akustik modellerin eğitiminde Saklı Markov Modelleri Yazılımı (HTK) kullanıldı [10].

Sınamalar iki farklı test verisi kullanılarak yapıldı. Kullanılan ilk test verisi radyoloji raporlarının okunduğu ses kayıtlarından oluşmaktadır. Bu kayıtların özelliği belirli cümle kalıplarının sıklıkla tekrar edilmesi ve kullanılan farklı kelime sayısının daha genel bir alana göre sınırlı olmasıdır. Ayrıca bu cümlelerde radyoloji ve tıp alanına özgü terimler sıkça kulanılmaktadır. Bu özelliklerinden dolayı, bu veri bildirinin devamında "özel test verisi" olarak isimlendirilecektir.

Radyoloji raporlarına özel dil modeli oluşturmak için yine radyoloji alanında toplanmış 874 farklı cümle kullanıldı. Bu metin verisi ile radyoloji alanına özel ikili dil modeli (bi-gram) oluşturuldu. Oluşturulan dil modeli Bölüm 3'de "radyoloji (özel) dil modeli" olarak anılacaktır. Bildirinin 4. bölümündeki "özel test verisi" tanıma sonuçları bu ikili dil modeli kullanılarak elde edilmiştir.

Sınamalarda kullandığımız ikinci test verisi bir gazetenin internet sitesinden spor, siyaset, yazarlar, güncel-magazin ve teknoloji gibi farklı haber kategorilerinde rastgele toplanmış metinlerden oluşmaktadır. Bu metinleri kullanarak kaydedilmiş ses verileri bildirinin devamında "genel test verisi" olarak isimlendirilecektir. "Genel test verisinde" kullanılabilecek farklı kelimelerin sayısı "özel test verisine" göre daha fazladır. Bu nedenle bu test verisinde Türkçe LVCSR uygulamalarında karşılaşılan kapsama problemiyle karşılaşılacaktır. Bu yüzden "genel test verisiyle" elde edilen tanıma sonuçlarına, yazılımın LVCSR uygulamalarında göstereceği performans olarak bakılabilir.

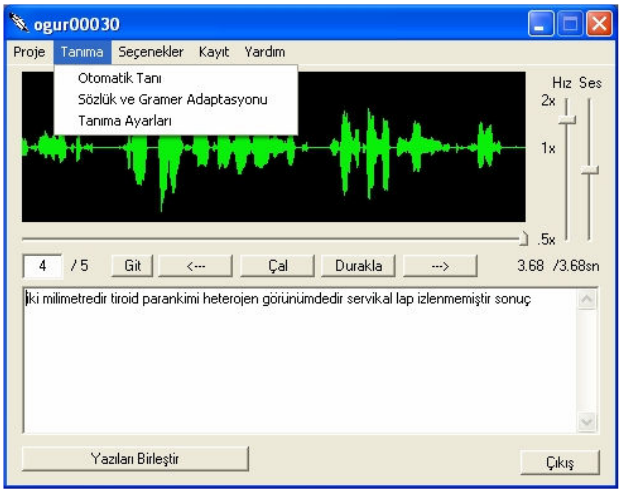
LVCSR uygulamalarında kullanılacak dil modelinin eğitimi için Sabancı Üniversitesinde toplanmış yaklaşık 5.5 milyon cümleden oluşan metin verisi kullanıldı [6]. Metin verisinde en sık geçen kelimeler listelendi ve bu listede en sık 500-1000 arasındaki kelimeler bulundu. Bu 500 kelimenin de içinde geçtiği 50k cümle büyük metin verisinden seçildi. Dil modeli adaptasyonunu genel dil modeli için de kullanılabilir kılmak için, büyük metin verisinin tamamı yerine seçilen 50k cümle ikili dil modeli eğitiminde kullanıldı. Bu dil modeli seçilmiş metin verisinde en sık geçen 20k kelime için oluşturuldu. Bu model Bölüm 3'de "genel dil modeli" olarak anılacaktır.

Bahsedilen "özel" ve "genel" dil modelleri kullanıcının isteğine göre adapte edilebilmektedir. Kullanıcı sıklıkla kullandığı cümleleri ya da kelimeleri daha önce oluşturulmuş dil modellerinin üstüne ekleyebilmektedir. Bu şekilde Türkçe LVCSR uygulamalarında karşılaşılan kapsama problemine pratik bir çözüm sunulmuştur. Ayrıca daha önce oluşturulmuş dil modelleri kullanıcının işine yaramadığı durumlarda, kullanıcı istediği alanda kendisine özel dil modeli oluşturup ses tanıma uygulamasını bu dil modeliyle kullanabilmektedir. Bu yönleriyle yazılım, tanıma yapılmak istenen alana kolaylıkla uyarlanabilmektedir. Kullanıcının dikte işini kolaylaştıracak bu tür uygulamalar, Bölüm 3'te yazılımın arayüzünün detayları anlatılırken ayrıntılı bir şekilde gösterilecektir.

3. Yazılımın Kullanımı ve Arayüzü

3.1. Ana arayüz

Geliştirilen yazılımın ana arayüzü Şekil 1'de görülebilir:

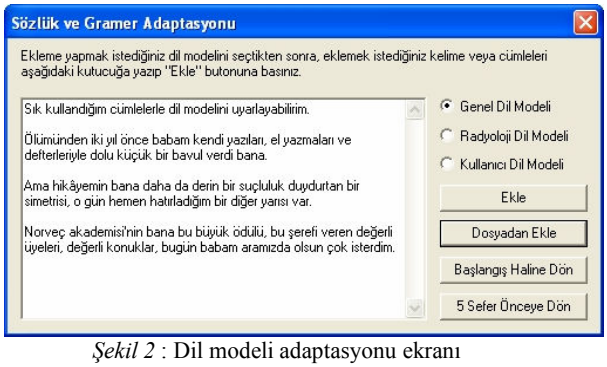


Şekil 1 : Yazılımın ana arayüzü

Ses tanıma işlemi tamamlandıktan sonra tanınan metin kullanıcının istediği uzunlukta bölütler halinde ekranda yansıtılmaktadır. Kullanıcı hipotez ve referans kelimeler arasındaki karşılaştırmayı, arayüz sayesinde ses dosyasının tanınan kelimelere denk gelen kısmını dinleyerek kolaylıkla yapabilmektedir. Programda bu karşılaştırmayı kolaylaştırmak için ses dosyalarını farklı çalma hızlarında dinleme olanağı da sunulmuştur.

Ana arayüzdeki "Tanıma" menüsünden "Sözlük ve Gramar Adaptasyonu" seçilirse, mevcut dil modelleri istenilen kelime ya da cümlelerle adapte edilebilmektedir.

3.2. Dil Modeli Adaptasyonu



Şekil 2 : Dil modeli adaptasyonu ekranı

Yazılıma, Türkçedeki kapsama problemini pratik bir şekilde çözmek için dil modeli adaptasyonu seçeneği eklenmiştir. Kullanıcı bu seçenek ile kelime dağarcığını ve dil modelini kendi kullanımına uygun bir şekilde genişletebilmektedir. Genişletme işlemini "Dosyadan Ekle" seçeneğini ile daha önceden oluşturduğu metin dosyasını seçerek ya da yukarıdaki örnekte görüldüğü gibi istediği cümleleri ekrana yazarak yapabilmektedir.

Kullanıcı tarafından eklenen bu cümleler uyarlama yapılmak istenen model için daha önceden hazırlanmış metinlerin sonuna eklenmekte ve ikili dil modeli eklenen cümlelerle birlikte yeniden oluşturulmaktadır. Çalışmanın ileriki aşamalarında dil modeli uyarlaması, eklenen cümlelerle oluşturulan dil modeli ile uyarlama yapılmak istenen dil modelinin aradeğerlemesi ile yapılacaktır. Aradeğerleme sonucunda en iyi başarımı elde etmek için değişik aradeğerleme katsayıları da denenecektir.

Kullanıcı daha önceden eğitilmiş 3 ayrı dil modelinin üstüne ekleme ve uyarlama yapabilmektedir:

Genel Dil Modeli:

Bu dil modeli Bölüm 2'de bahsedilen LVCSR uygulamalarında kullanılmak için hazırlanmış ikili dil modelidir. Bu model kullanılarak genel test verisinde elde edilen tanıma sonuçları bildirinin 4. bölümünde verilecektir.

Radyoloji (Özel) Dil Modeli:

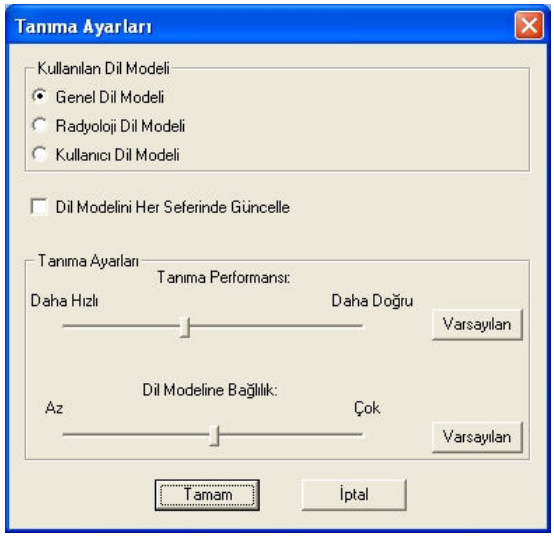
Bu dil modeli Bölüm 2'de belirtildiği gibi radyoloji alanına özgüdür. Bu model kullanılarak özel test verisinde elde edilen tanıma sonuçları bildirinin 4. bölümünde verilecektir.

Kullanıcı Dil Modeli:

Bu model önceden eğitilmiş modellerden bağımsız olarak sadece kullanıcının eklediği metinlerle oluşturulmuştur. Tanıma sözlüğünde eklenen metinlerdeki kelimeler vardır. Bu model sayesinde kullanıcı istediği alanda bir dil modeli oluşturabilmekte ve yazılımı özel (radyoloji) ve genel dil modellerinin dışındaki bir alanda da kullanabilmektedir.

Programa verilen ses dosyasının tanıması otomatik olarak son ekleme ve uyarlama yapılan dil modeli kullanılarak yapılır. Kullanıcı tanıma sırasında kullanılmasını istediği dil modelini "Tanıma Ayarları" ekranından seçebilmektedir.

3.3. Tanıma Ayarları



Şekil 3 : Tanıma ayarları ekranı

"Tanıma Ayarları" ekranında hem tanıma sırasında kullanılmak istenen dil modeli hem de tanıma performansını etkileyecek diğer parametreler seçilebilmektedir. Arayüzde görülen "Tanıma Performansı" ve "Dil Modeline Bağlılık" seçenekleri sırasıyla budama eşiği (beam searching threshold) ve gramar çarpanını (grammar scale factor) belirlemektedir. Bu seçenekler sayesinde kullanıcı daha hızlı ya da daha doğru tanıma arasında isteğine uygun bir seçim yapabilmektedir.

4. Deneyler ve Sonuçlar

Yazılımın tanıma performansı Bölüm 2'de belirtilen iki farklı test verisi kullanılarak ölçüldü: özel ve genel. Özel test verisi 10 farklı kişi tarafından kaydedilmiş yaklaşık 2 saatlik radyoloji raporlarından oluşmaktadır. Yazılımın LVCSR uygulamalarındaki performansını ölçmek için spor, siyaset, yazarlar, güncel-magazin ve teknoloji haberlerinden oluşmuş 100 farklı cümle kullanıldı. 8 farklı kişi tarafından bu cümleler kullanılarak kaydedilmiş yaklaşık yarım saatlik ses kaydı genel test verisini oluşturmaktadır. Testte kullanılan verilerin özellikleri ve dil modellerinin oluşturulması sırasında kullanılan kelimelerle elde edilen kapsama oranları aşağıdaki tabloda görülebilir:

	Test verisi miktarı	Farklı kelime sayısı	% Kapsama oranı
Özel	2 saat	559	96.37
Genel	0.5 saat	837	77.80

Tablo 1: Özel ve genel test verisi özellikleri

Tabloda görüldüğü gibi genel test verisindeki kapsama oranı özel test verisine göre daha düşüktür. Kapsama oranlarındaki bu fark iki test verisinin farklı özelliklerinden kaynaklanmaktadır. Özel test verisi kısıtlı bir alanda toplanmış cümlelerden oluşurken genel test verisi beş farklı haber kategorisinde toplanmış cümlelerden oluşturulmuştur.

Özel ve genel test verileriyle elde edilen tanıma sonuçları aşağıdaki tablolarda değişik budama eşiği değerleri için görülebilir. Tablolarda verilen tanıma hızı tanınan ses kayıtlarının toplam uzunluğunun, toplam tanıma süresine bölünmesi ile elde edilmiştir. Bütün tanıma deneyleri çift çekirdekli 1.83Ghz işlemcili 1024Mb bellekli dizüstü bilgisayarda yapılmıştır.

Budama eşiği	%Kelime doğruluğu	%Kelime kesinliği	%Harf kesinliği	Tanıma hızı
80	74.13	70.32	76.28	12.21
100	86.16	83.73	88.41	4.43
150	91.78	89.48	94.38	1.25
200	92.58	90.29	95.37	0.69

Tablo 2: Özel test verisi tanıma sonuçları

Budama eşiği	%Kelime doğruluğu	%Kelime kesinliği	%Harf kesinliği	Tanıma hızı
70	42.57	35.00	71.28	1.11
80	48.07	43.13	76.42	0.42
100	49.20	44.36	77.95	0.20
125	49.53	44.78	78.15	0.11

Tablo 3: Genel test verisi tanıma sonuçları

Tablolarda görüldüğü gibi yazılımın özel test verisindeki (radyoloji) tanıma performansı bu alandaki cümlelerin belli kalıpları olması ve bu kalıpların dil modeli tarafından öğrenilebilmesi nedeniyle oldukça başarılıdır. Radyoloji gibi sınırlı bir alanda yaklaşık %90 kelime tanıma peformansı tanınan ses kaydının uzunluğundan daha kısa bir sürede elde edilebilmektedir. Genel test verisi tanıma performansı ve hızı özel test verisine oranla daha düşüktür. Genel test verisinden elde edilen tanıma performansına bakılarak, yazılımın herhangi bir konuda söylenen her iki kelimeden yaklaşık bir tanesini genel dil modeli ile tanıyabileceği söylenebilir.

Her iki test verisinde elde edilen başarılı harf tanıma oranları, Türkçenin eklemeli yapısından kaynaklanmaktadır. Tanınan kelimelerin ek veya köklerinde yapılacak ufak hatalar tüm kelimenin yanlış tanınması olarak sayıldığından, harf tanıma oranları yazılımın gerçek tanıma performansı hakkında önemli bilgi içermektedir. Özellikle genel test verisindeki yüksek harf tanıma oranları, sözcüklerin ek veya köklerinde yapılmış ufak hatalar giderildiğinde Tablo 3'de görülen kelime tanıma oranlarından daha yüksek bir performansın elde edilebileceğini ortaya koymaktadır.

Ayrıca dil modeli adaptasyonu seçeneği ile kullanıcı istediği cümle kalıplarını ya da kelimeleri dil modeline ekleyerek tanıma performansını arttırma şansına sahiptir. Bu yönleriyle çalışmadan elde edilen sonuçlar Türkçe için hem sınırlı hem de geniş dağarcıklı ses tanıma uygulamalarının geliştirilmesi konusunda ümit vericidir.

5. Sonuçlar ve Gelecek Çalışmalar

Bu bildiride dil modeli kullanıcının isteğine göre uyarlanabilen bir ses tanıma yazılımının tanıtımı yapılmıştır. Bu yazılımın performansı radyoloji-özel ve geniş dağarcıklı-genel test verileri kullanılarak ölçülmüştür. Yazılımın radyoloji gibi sınırlı alanlarda hem hız hem de tanıma performansı olarak başarılı olabileceği görülmüştür. Genel test verisinde ise hız ve tanıma performansının arttırılması gerekmektedir. Her iki test verisinde elde edilen yüksek harf tanıma oranları, Türkçenin eklemeli yapısından kaynaklanan problemler giderildiğinde elde edilecek performans hakkında umut verici sonuçlar içermektedir.

Gelecek çalışma olarak geniş dağarcıklı ses tanıma uygulamalarındaki başarım arttırılmaya çalışılacaktır. Ayrıca dil modeli uyarlaması için dil modeli aradeğerlemesi gibi yöntemler yazılıma eklenecektir.

6. Kaynakça

[1] Osman Buyuk S*ub-word Language Modeling for Turkish Speech Recognition,* M.S. Thesis, Sabanci University, 2005.

[2] Ebru Arisoy, *Turkish dictation system for radiology and broadcast news applications,* M.S. Thesis, Bogazici University, 2004.

[3] Kenan Carkı, Petra Geutner, and Tanja Schultz, "Turkish LVCSR: Towards better speech recognition for agglutinative languages," *ICASSP 2000*, Istanbul, Turkey, 2000, vol. 3, pp. 1563–1566.

[4] Erhan Mengusoglu and Olivier Deroo, "Turkish LVCSR: Database preparation and language modeling for an agglutinative language," *ICASSP 2001, Student Forum*, Salt-Lake City, 2001.

[5] Ebru Arısoy and Levent M. Arslan, "Turkish dictation system for broadcast news applications," *EUSIPCO 2005*, Antalya, Turkey, 2005.

[6] Hakan Erdogan, Osman Buyuk, and Kemal Oflazer, "Incorporating language constraints in sub-word based speech recognition," *ASRU 2005*, Cancun, Mexico, 2005.

[7] Kadri Hacioglu, Brian Pellom, Tolga Ciloglu, Ozlem Ozturk, Mikko Kurimo, and Mathias Creutz, "On lexicon creation for Turkish LVCSR," *EUROSPEECH 2003*, Geneva, Switzerland, 2003, pp. 1165–1168.

[8] Ebru Arisoy and Murat Saraclar, "Lattice extension and Rescoring Based Approaches for LVCSR of Turkish", *INTERSPEECH 2006*, Pittsburg, USA, 2006

[9] Ozgur Salor, Bryan Pellom, Tolga Ciloglu, Kadri Hacioglu, Mubeccel Demirekler, "On Developing New Text and Audio Corpora and Speech Recognition Tools for the Turkish Language", *ICSLP*, 2002.

[10] S. Young, D. Ollason, V. Valtchev and P. Woodland, *The HTK Book (for HTK Version 3.2),* Entropic Cambridge Research Laboratory, 2002.