KONUŞMA TANIMADA KARMA DİL BİRİMLERİ KULLANIMI VE DİL KISITLARININ GERÇEKLENMESİ

USING HYBRID LEXICON UNITS AND INCORPORATING LANGUAGE CONSTRAINTS IN SPEECH RECOGNITION

Osman Büyük, Hakan Erdoğan, Kemal Oflazer

Mühendislik ve Doğa Bilimleri Fakültesi
Sabancı Üniversitesi, Orhanlı Tuzla, 34956, İstanbul

buyuk@su.sabanciuniv.edu.tr, {haerdogan,oflazer}@sabanciuniv.edu.tr

Özetçe

Eklemeli diller için geniş dağarcıklı konuşma tanıma uygulamalarında bütün sözcükleri kapsama ile ilgili sorunlar çıkmaktadır. Sabit bir sözcük listesi yeterli olmadığından sözcük-altı dil birimlerinin kullanımı gerekmektedir. Karma dil birimleri kullanımı ve bu birimler üzerinden dil modeli geliştirilmesi teorik olarak kapsama sorununa bir çözüm sağlar. Fakat bu durumlarda da kısa birimler kullanımı nedeniyle akustik karışıklık sorunları ve dil modelinin uzun geçmiş kullanamama sorunları karşımıza çıkar. Biz bu çalışmada karma dil birimlerinin seçimi ve bunların üzerinden geliştirilecek dil modeli içine dilden gelen kısıtların da dahil edilmesi üzerinde duracağız. Hem istatistiksel dil modelini hem de dilden gelen bazı kısıtlamaları "ağırlıklı sonlu durum makinesi" ile ifade ederek birleştirebilir ve sonuçta yeni ve daha iyi bir dil modeli elde edebiliriz. Bu çalışmada bu birleşik dil modelinin başarımını inceliyoruz. Dilden gelen kısıtlamaları ifade eden ağırlıklı sonlu durum makinesi, 2151 sözcüklü bir test verisinde sözcük hata oranında oransal olarak %3 indirim sağlamaktadır.

Abstract

We face problems in large vocabulary continuous speech recognition for agglutinative languages, due to lack of coverage of all possible words. Since it is not enough to have a finite full-word lexicon, we may use sub-word units for recognition. Using sub-word lexicon units and developing language models based on these units solves the coverage problem. However, this results in increased acoustic confusability and shorter effective language model history length. We introduce new ways to choose lexicon units and we incorporate linguistic constraints into a statistical language model developed with the new units. We represent both the statistical language model and linguistic constraints as weighted finite state machines (WFSM) and combine them to obtain a novel language model. We study the performance of the new language model and show that it achieves 3% relative reduction in word error rate when used in recognizing a test-set of 2151 words.

1. Giriş

Modern konuşma tanıma sistemlerinde akustik modelleme için saklı Markov modelleri (hidden Markov models – HMM) kullanılır. Dili modellemek için de tanımanın geniş dağarcıklı olup olmadığına göre değişebilen istatistiksel ya da kural temelli gramerler kullanılır. Eğer konuşma tanıma uygulaması sınırlı bir gramer kullanımını mümkün kılıyorsa, elle yazılan bu tür gramerler kullanıldığında başarım oldukça yüksek olur. Fakat bu gramere uymayan cümleler için sistem tanımayı sağlayamayacağından, başarısız sayılacaktır. Sesli diyalog sistemlerinde eğer kullanıcı sorularla iyi bir şekilde yönlendirilebilirse dar gramer temelli tasarım iyi çalışabilir. Fakat yine de en uygun çözüm geniş dağarcıklı (yani sınırsız) konuşmaya izin vererek konuşma tanıma başarımını arttırmaktır. Ayrıca dikte ve haber yayınlarını tanıma gibi uygulamalarda her zaman geniş dağarcıklı tanıma gereklidir.

Geniş dağarcıklı tanıma için eğitilebilir N-gram dil modelleri İngilizce için iyi sonuçlar vermektedir. Bu modeller ile tanıma başarımı İngilizce dikte için %95'ler seviyesinde olabilmektedir. Türkçe gibi eklemeli dillerde ise sözcük listesi çok çok büyük olduğundan ve önceden belirlenen listeler çoğu zaman yetersiz kaldığından direk sözcük birimleri ile tanıma başarısız olmaktadır [1][2][3][4]. Bu durumda sözcük birimleri yerine sözcük-altı birimlerin kullanımı bir çözüm olarak ortaya çıkmaktadır [1][2][3][4][5][6]. Bu alt birimler, sözcüğün gövdesi ve ekleri (yani morfolojik birimler) veya heceler olabilir [5][6].

Ne var ki, sözcük-altı birimlerin konuşma tanımada kullanımı da her zaman çok iyi bir başarım iyileştirmesi gerçekleştirememektedir [3][7]. Bunun sebeplerinden biri alt-birimlerin genelde çok kısa olmaları ve bu kısa alt birimler arasındaki akustik karışıklıkların artmasıdır. Diğer bir sebep de dil modelinin etkin olarak daha kısa bir geçmişe bakma durumunda kalması ve geniş geçmişi görememesidir. Bu etkenler göz önüne alındığında karma birimlerin kullanımı gerekmektedir. Bu birimlerin nasıl seçilmesi gerektiği çok önemlidir. Burada ana amaç dili mümkün olduğunca kapsarken başarımı olumsuz etkilememek için çok kısa veya gereksiz alt birim kullanmamaktır. Bu bildiride karma birim listesi oluşturma ile ilgili birtakım fikirler teklif ediyoruz.

Bir diğer sorun da ses tanıma sisteminde sözcük-altı birimler kullanıldığında dil kurallarına aykırı sözcüklerin kodçözücü tarafından çıkarılabilme olasılığının bulunmasıdır. Biz bu bildiride bu olasılıkları engellemek için ağırlıklı sonlu durumlu makinelerinin (weighted finite state machine – WFSM) kullanılmasını öneriyoruz. Dilden gelen kısıtları WFSM ile modelleyerek dil modelini kısıtlamak yeni bir yaklaşım tarzıdır ve bu yolla hem dil modeli boyutu düşürülüp hem de daha yüksek başarım elde edilebilir.

Bu çalışma şu şekilde düzenlenmiştir. Bölüm 2'de genel olarak N-gram dil modelleme anlatılacaktır. 3. bölümde birleşik dil modelinin WFSM ile gerçeklenmesinden bahsedeceğiz. 4. bölümde yeni değerlendirme ölçütlerimizi tanıtacağız. Yaptığımız deneyleri 5. bölümde anlatacağız. Bildirinin sonuçlarını ve geleceğe dönük planlarımızı son bölümde bulabilirsiniz.

2. İstatistiksel Dil Modelleme

Dil modelleri bir dildeki cümlelerin olasılığının bulunabilmesi için geliştirilen modellerdir. Cümlelerin sözcük dizilerinden oluştuğu varsayılırsa bu olasılık şöyle yazılabilir [8].

$$P(W) = P(w_1, w_2, ..., w_n) = \prod_{i=1}^{n} P(w_i \mid w_1, ..., w_{i-1}).$$

Burada W cümleyi, w_i ise sözcükleri gösterir. N-gram modelleri denklemin sağındaki olasılıkları sadece geçmişteki N-1 terimi kullanacak şekilde yakınsar. Yani,

$$P(W) \approx \prod_{i=1}^{n} P(w_i \mid w_{i-N+1}, ..., w_{i-1})$$

şekline getirir. Her bir terimin hesaplanabilmesi için en iyi olabilirlik (maximum likelihood) kuralı ile elimizdeki eğitim için toplanmış metin verisini kullanarak her bir N-gram'ın görülme oranları hesaplanır. Eğitim verisinde görülemeyen N-gram'ların sıfır olasılık almamaları için ise birtakım yumuşatma algoritmaları geliştirilmiştir [9].

Giriş bölümünde de belirttiğimiz gibi Türkçe için kelime sayısı çok yüksek, adeta sınırsız olabildiğinden dil modelinde temel birim olarak kelimelerin kullanımı mümkün değildir. Bu durumda daha alt birimler kullanımı gerekli ve faydalı olacaktır.

Bir kelimenin alt birimlere ayrılması birden çok şekilde gerçeklenebilir. Örneğin ilk akla gelebilen hecelere ayırmaktır. Aşağıda çeşitli parçalara ayırma metotları listelenmiştir: (1) Sözcüğü bir bütün olarak alma, (2) Gövde + ek-dizisi, (3) Gövde + ek1 + ek2 + ... , (4) Hecelere ayırma. Biz bu çalışmada bir sözcüğü parçalara ayırırken (veya tersine parçalardan oluştururken) sırasıyla 1, 2 ve 4. seçenekleri değerlendireceğiz. Sözcüğü bir bütün olarak dağarcık listemizde bulamazsak, onu iki parçaya ayırmaya çalışacağız. Eğer o da mümkün olmazsa hecelerine ayıracağız. 3. seçeneği kullanmaya, akustik karışıklığı arttıracağını düşündüğümüzden ve 2. seçenek tarafından parçalanamayan kelimelerin büyük ihtimalle 3. seçenek tarafından da parçalanamayacağını düşündüğümüzden gerek görmüyoruz.

Bu tür bir ayrımdan sonra dağarcık listemizdeki birim türleri şu şekillerde sınıflandırılabilir:

1. Gövdeler: tam sözcük ya da ilk yarı-sözcük olarak kullanılırlar (örneğin: ev, sokak, duygu, gerilim, v.s.)
2. Ek-dizileri: ikinci yarı-sözcük olarak kullanılırlar (örneğin: -ler, -lar, -lerde, -imizin, -imizdekiler, -ıldığında v.s.)
3. Heceler: (örneğin: <a>, <e>, <de>, <bak>, <kır>, <trak>,<ler#>, v.s.)

Bazı heceler, gövde olarak da düşünülebilir, örneğin "bak". Bu durumlarda hece ve kök "bak" arasındaki ayırımı sağlayabilmek için dağarcık listemizde hece birimlerini "<" ve ">" sembolleri arasında gösterir ve ayırt etmeyi sağlarız. Ayrıca sözcük sonundaki heceleri de # işareti ile ayrıştırırız. Aynı şekilde dağarcığımızda ek dizilerinin başına "-" koyarak ek dizileri ile heceler ve gövdeler arasındaki olası karışıklığı da engelleriz.

3. Geliştirilen Birleşik Dil Modeli

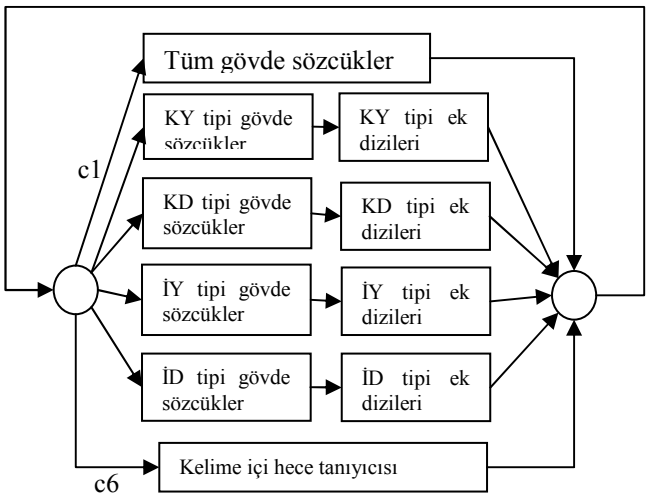
N-gram dil modelleri hazır yazılımlar ile bir WFSM haline dönüştürülebilirler [10][11]. Sonuç olarak karmaşık ve büyük bir WFSM ortaya çıkar. Bu WFSM içerisindeki ağırlıklar istatistiksel olarak eğitilmiştir. Buradaki ağırlıklar birimlerin belli bir geçmişi izleme olasılıklarından oluşur. Bu olasılıkların güvenilir olarak eğitilmesi için milyonlarca hatta bir milyara yakın birimden (çoğunlukla kelime) oluşan metin verisi gereklidir. Bu verilerin de çoğu zaman temizlenmesi gerekir ve temizleme işlemi mükemmel olmadığından istatistiksel eğitim içerisinde az da olsa gürültü bulunur.

Özellikle sözcük-altı birimlerin kullanıldığı durumlarda ve sadece istatistiksel dil modeli kullanıldığında konuşma kodçözücüsü dil açısından kesinlikle mümkün olmayan alternatifleri de çıkarabilir. Örneğin yukarıdaki belirttiğimiz karma birimler kullanıldığında kodçözücü şöyle bir cümle çıkarabilir:

<a> <de> <la> <yıd> ev -inde -ler <a> <vi> <za> bul –ındırmaz.

Bu cümlenin orjinalinin "Adelaide evinde avize bulundurmaz." olduğunu varsayalım. Kodçözücü çıktısında görüldüğü gibi dilbilgisi bakımından olası olmayan durumlar ortaya çıkmıştır. Bu örnekte görülen sorunları listelersek: (1) İki ek dizisi arka arkaya gelebilir (-inde -ler gibi). (2) Gövdeyi takip eden ek dizisi sesli uyumuna uymayabilir (bul –ındırmaz gibi). (3) Kelime dağarcıkta olsa dahi (avize) onun yerine akustik olarak benzer hece dizisi tercih edilebilir.

Aslında iyi eğitilmiş bir istatistiksel dil modelinin bu tür sorunlu durumları azaltması beklenir ama yine de bu tür durumların olasılığı sıfırlanmamıştır. Yukarıda listelediğimiz türden hataları engellemek için biz yeni bir kural tabanlı WFSM kullanmayı öneriyoruz.
Bu WFSM, yukarıdaki sorunları engellemek için doğru yarı-sözcük dizilerine izin verecek şekilde tasarlanmıştır. Tasarlanan WFSM'i kavramsal olarak *Şekil 1*'de görebilirsiniz. Bu WFSM'in ana amacı Türkçe'deki sesli uyumunu zorlamak ve doğru yarı-sözcük sıralamasını zorlamaktır.



Şekil 1: Teklif edilen dil kurallarına uygun kural tabanlı WFSM'in kavramsal gösterimi

Yukarıdaki şekilde her bir branş için kullanılacak ağırlıklar c1,...,c6 şeklinde gösterilmiştir. Buradaki ağırlıklar ile oynanarak dil modelinin hece modelini çok mecbur kalmadıkça tercih etmesi engellenebilir. Böylece yukarıda bahsettiğimiz 3. tip sorun da çok fazla görülmeyecektir. c6 katsayısı sıfır alındığında elimizde sadece yarı-sözcüklerden oluşan bir dağarcık ve dil modeli yeterli olur. Bu çalışmadaki sınamalarımızda biz yarı-sözcük tabanlı sonuçları sunuyoruz.

Şekil 1'de görülen WFSM istatistiksel N-gram WFSM ile birleştirilecek (fsm-compose işlemi ile) ve sonuçta hem yazı verisinden eğitilmiş hem de dil kurallarına uygun bir dil modeli elde edilmiş olacaktır. Yani efektif olarak bir birim dizisinin olabilirliği N-gram dil modelinden hesaplanacak ama kural tabanlı WFSM tarafından kabul edilmeyen birim dizilerine hiçbir zaman izin verilmeyecektir.

4. Değerlendirme Ölçütleri

Ses tanıma başarımını ölçmek için sözcük hata oranları (word-error-rate – WER) bilinen bir ölçüm biçimi olmakla birlikte Türkçe gibi eklemeli dillerde başarım oranını ölçmek için sözcük-altı birimlere dayanan yeni ölçütlere ihtiyaç duyulmaktadır. Bir kod-çözücü genellikle kelimenin gövdesini doğru tanırken, ek kısımlarında hata yapabilmektedir. Sözcük hata oranını tanıma başarımını ölçmede bir ölçüt olarak kullandığımızda, ek-dizisindeki ufak bir hatadan dolayı tüm sözcük hatalı sayılmaktadır. Bununla beraber sözcüğün gövdesi doğru olarak tanındığından, sadece ek kısımda yapılan hatayı tam bir hata saymak yerine, bir doğru tanıma ve bir yerine geçme hatası olarak saymak daha doğru bir değerlendirme şekli olacaktır. Bir başka seçenek ise tanınan kelimenin ek kısmını atarak sadece gövde kısmını kullanıp hata oranlarını hesaplamaktır. Bu değerlendirme yöntemi, ses tanıma sisteminin kelimenin esas kısmını (gövdesini) doğru tanımasıyla ilgili bilgiyi bize verecektir.

Böylece ses tanıma sistemimizin başarım oranını hesaplarken üç değişik ölçüt kullandık:

1. **WER**: Sözcük hata oranı (word-error-rate). Kelime dizilerini tanınan sözcük-altı birimlerden oluşturduk (Bu oluşturma şekli genellikle belirsiz değildir çünkü gövde + ekler ve # sembolüyle biten hece dizileri bir kelimeyi oluşturmaktadır. Sadece istatistik kullanan dil modellerinde izin verilmeyen sözcük-altı dizilerine de sahip olabiliriz. Bu durumlarda birleştirilemeyen sözcük-altı birimler oldukları gibi kalırken, birleşebilenlerse sözcük haline getirilmektedir.)
2. **HWER**: Yarı-sözcük tanıma oranı (half-word-error-rate). Birleştirilen sözcükleri gövde + ek-dizisi şeklinde tekrar iki parçaya ayırdık. Bu ayırma işlemini morfoloji deneyinde kullandığımız çözümleyiciyi kullanarak ayrıca tek karakterden oluşan eklere de izin vererek elde ettik. Bu şekilde elde edilen yarı-sözcük birimleri arasındaki hata oranlarını hesapladık.
3. **STER**: Gövde tanıma oranı (stem-error-rate). Birleştirilen sözcükleri iki parçaya ayırdıktan sonra, ikinci parçayı sildik. Böylece sadece gövdeler arasındaki hata oranlarını hesapladık.

Deneylerimizde hata oranlarını yukarıda bahsedilen ölçütleri kullanarak hesapladık.

5. Deneyler ve Sonuçlar

Türkçe için yaptığımız geniş dağarcıklı ses tanıma deneyi için Sabancı Üniversitesi ve ODTÜ'de [12] toplanan ses verilerini kullandık. Sabancı Üniversitesinde toplanan verilerde 367 farklı kişiden fonetik olarak dengeli yaklaşık 100 cümle okumalarını istedik. Ses verisi toplanan 367 kişi arasında yaklaşık eşit sayıda kadın ve erkek olmasına dikkat ettik. Kullandığımız bütün ses verilerini 16 KHz de örneklerken nicemlemek için 16 bit kullandık. Deneylerimizi yaklaşık 37 saatlik ses verisi üzerinde yaptık. 37 saatlik ses verisinin 34 saatini Saklı Markov Modellerinin eğitimi için ayırırken geriye kalan 3 saatlik ses verisini elde edilen eğitim modellerinin sınanması için kullandık. Sınama verimiz 16 farklı kişiden toplanan ve değişik gazetelerden alınmış 88 spor haberi cümlesini içermektedir.

Türkçe gibi eklemeli dillerde yapılacak ses tanıma sistemlerinde büyük bir öneme sahip olan N-gram dil modelinin eğitimi için değişik gazetelerin sitelerinden ve İnternet'te bulunan elektronik kitaplardan faydalandık. Bu yöntemlerle elde ettiğimiz metin verileri dört ana gruba ayrılabilir: spor haberleri, güncel haberler, yazarlar ve elektronik kitaplar. Böylece dil modelinin eğitimi sırasında toplam 5.556.449 farklı cümle kullanırken, bu metin verisinden toplam 1.170.526 adet farklı kelime elde ettik. Elde edilen farklı kelime sayısının fazlalığı Türkçe için kelime tabanlı ses tanıma sistemlerindeki zorluğu da ortaya koymaktadır.

Deneyler sırasında sözcük veya sözcük-altı 3 farklı dağarcık birimi kullanılmıştır: Kelime, hece ve yarı-sözcükler. Yarı-sözcükler deneyinde kelimelerin morfolojik kısımları Kemal Oflazer'in Türkçe için geliştirilmiş morfolojik çözümleyicisinden elde edilmiştir [13]. Türkçe'de bir kelimenin birden fazla şekilde morfolojik kısımlarına ayrılması mümkün olabilmektedir. Bu gibi durumlarda bütün kelimelere uygulanan ve kelime için olası ek ve kök parçalarından birisini seçen genel bir işlem geliştirilmiştir. Bu

işlemde kelime için mümkün olan en uzun kök parçası alınırken ek parçasınında birden fazla harften oluşmasına dikkat edilmiştir. Yarı-sözcük deneyinde en fazla geçen 10000 gövde ve 3000 ek-dizisi kullanılmıştır. Bir kelime eğer en fazla geçen 10000 gövde arasında bulunursa kelime bir bütün olarak alınırken, bu durum olası değilse kelime en fazla geçen gövde ve ekleri kullanarak "gövde+ek-dizisi" şeklinde ayrılmaya çalışılmıştır.

Tanıma hata oranları her deney için elde edilen bi-gram dil modeli ile akustik modelin birleştirilmesi ile elde edilmiştir. Sözcük tabanlı sistem ile daha iyi bir karşılaştırma yapabilmek için yarı-sözcük ve hece tabanlı deneylerde, kelime tanıma oranları da elde edilmiştir. Bu deneyler için elde edilen hata oranları aşağıdaki tablolarda görülebilir:

Dağarcık tipi	Doğru cümle %	WER	HWER	STER
Sözcük	11.93	52.95	42.18	43.21
Hece	2.85	68.80	64.63	62.71
Yarı-sözcük	11.99	45.11	40.12	36.30

Tablo 1: 16 kişinin verisinde bi-gram dil modeli kullanarak dağarcık tipine göre karşılaştırılabilir hata oranları

Tanıma oranlarında görüldüğü gibi bi-gram dil modeli ile en iyi sonuçlar yarı-sözcük tabanlı deneyde elde edilmiştir. Bu başarım nedeniyle kural tabanlı WFSM sadece bu dağarcık tipi ile uygulanmıştır. Bir alt sınama kümesi olarak 3 farklı konuşmacı için HTK'de lattice'ler oluşturulmuştur. Tri-gram dil modeli eğitimi de metin verisindeki 2 milyon cümle kullanılarak yapılmıştır ve bu dil modelinin WFSM'inin lattice'ler ile kural bazlı WFSM sonrasında kompozisyonu alınmıştır. Bu deneyden elde edilen hata oranları aşağıdaki tablo ile özetlenebilir:

Yarı-sözcük dağarcığı ile kullanılan dil modeli	Doğru cümle %	WER	HWER	STER
Bi-gram	11.74	40.51	36.51	31.53
Kural-tabanlı WFSM	11.79	39.66	34.89	30.26
Tri-gram	19.70	33.06	30.19	25.83
Kural-tabanlı WFSM + Trigram	19.77	32.54	29.10	25.28
Lattice'den gelebilecek en iyi sonuç	31.44	21,25	19.20	15.14

Tablo 2: 3 kişinin verisi üzerinde dil modeli tipine göre karşılaştırılabilir hata oranları

Buradan da görüldüğü gibi tri-gram dil modeli bi-gram model üzerine yaklaşık %20 oranında hataları azaltmaktadır. Kural tabanlı WFSM uygulaması da tri-gram modelinin üzerine %3 oranında hata oranlarında indirim sağlamaktadır. Dolayısıyla böyle kural tabanlı ek bir dil modelinin faydaları bu sonuçlarda açıkça görülmektedir.

6. Sonuçlar ve Gelecek Çalışmalar

Bu çalışmada ağırlıklı sonlu durumlu makineler kullanarak Türkçe gibi eklemeli diller için geliştirilen yeni bir dil modeli tanıtılmıştır. Bu yeni dil modelinin geniş dağarcıklı konuşma tanıma için ümit vaat ettiğini düşünüyoruz ve gelecekte bu çerçevenin daha geliştirilerek kullanılabileceğini ummaktayız.

İlerde yeni dil kuralları bu çerçeveye eklenebilir. Gelecekte bu tür incelemelere devam edeceğiz.

7. Kaynakça

[1] Kadri Hacioglu, Bryan Pellom, Tolga Ciloglu, Ozlem Ozturk, Mikko Kurimo and Mathias Creutz, "Word splitting for Turkish LVCSR," *SIU,* 2003.

[2] Kadri Hacioglu, Bryan Pellom, Tolga Ciloglu, Ozlem Ozturk, Mikko Kurimo and Mathias Creutz, "On lexicon creation for Turkish LVCSR," *Eurospeech,* 2003.

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

[4] Helin Dutagaci and Levent M Arslan, "A comparison of four language models for large vocabulary Turkish speech recognition," *ICSLP,* 2002.

[5] Jeff Bilmes and Katrin Kirchhoff, "Factored language models and generalized parallel backoff," *Human Language Technology Conference,* 2003.

[6] Vesa Siivola, Teemu Hirsimaki, Mathias Creutz and Mikko Kurimo, "Unlimited vocabulary speech recognition based on morphs discovered in an unsupervised manner," *Eurospeech,* pp. 2293--2296, 2003.

[7] W Byrne, J Hajie, P Ircing, F Jelinek, S Khudanpur, P Krbec and J Psutka, "On large vocabulary continuous speech recognition of highly inflectional language - Czech," *Eurospeech,* 2001.

[8] Daniel Jurafsky and James H Martin, *Speech and language processing,* Prentice Hall, New Jersey, 2000.

[9] Stanley F Chen and Joshua Goodman, *An empirical study of smoothing techniques for language modeling,* Technical Report, no. 1098, Center for Research in Computing Technology, Harvard University, Aug, 1998.

[10] M Mohri and M Riley, "Weighted finite-state transducers in speech recognition (tutorial)," *ICSLP,* 2002.

[11] AT&T grm tools library, http://www.research.att.com/projects/mohri/grm .

[12] ODTÜ verisi Dr. Tolga Çiloğlu tarafından sağlanmıştır.

[13] K. Oflazer, "Two-level Description of Turkish Morphology," *Literary and Linguistic Computing, Vol.9 No.2*,1994.