Türkçe Konuşma Tanıma İçin Farklı Dil Modellerinin İncelenmesi
Investigation of Different Language Models for Turkish Speech Recognition

Ali Orkan Bayer, Tolga Çiloğlu†, Meltem Turhan Yöndem

Bilgisayar Mühendisliği Bölümü
†Elektrik ve Elektronik Mühendisliği Bölümü
Orta Doğu Teknik Üniversitesi, Ankara
orkan@ceng.metu.edu.tr, ciltolga@metu.edu.tr, mturhan@ceng.metu.edu.tr

Özetçe

Geniş dağarcıklı sürekli konuşma tanıma İngilizce gibi zengin biçimbilimsel özellikleri bulunmayan diller için yüksek başarılarla yapılabilmektedir. Halbuki, eklemeli dillerde başarı oranı çok düşüktür. Bunun en büyük nedeni diğer diller için kullanılan sözcükler üzerinden oluşturulan dil modellerinin eklemeli diller için yetersiz kalmasıdır. Bu çalışmada eklemeli dillerin yapısı göz önüne alınarak üç çeşit dil modeli incelenmiştir. Bu modellerden ikisi sözcük altı birimleri üzerinden çalışmaktadır. Bunlardan ilki, sadece sözcük gövdelerini dil modeli birimi olarak kullanırken, diğeri sözcük gövdeleri ile bunlardan sonra gelen ek dizilerini ayrı ayrı dil modeli birimi olarak kullanır. Üçüncü model ise önce sözcüklerin birlikte görülme olasılıklarına göre onları birer sınıfa yerleştirir ve daha sonra bu sınıfları model birimi olarak kullanır. Modellerin başarıları iki aşamalı tanıma yapılarak sınanmıştır; önce 2-çekirdekli modeller kullanılarak örgüler oluşturulmuş, daha sonra bu örgüler üzerinde 3-çekirdekli modeller denenmiştir. Bu çalışmanın sonucunda sözcük dağarcığı kapsama oranının, tanıma performansı üzerinde etkili olduğu görülmüştür. Bu nedenle, sözcük kapsamı bakımından üstün olan gövde ve ek dizileri üzerinden eğitilen modeller daha başarılı olmuştur. Ayrıca, bu modeller üzerinden tek aşamada yapılabilecek tanıma işleminin başarıyı daha da arttırabileceği düşünülmektedir.

Abstract

Large vocabulary continuous speech recognition can be performed with high accuracy for languages like English that do not have a rich morphological structure. However, the performance of these systems for agglutinative languages is very low. The major reason for that is, the language models that are built on the words do not perform well for agglutinative languages. In this study, three different language models that consider the structure of the agglutinative languages are investigated. Two of the models consider the subword units as the units of language modeling. The first one uses only the stem of the words as units, and the other one uses stems and endings of the words separately as the units. The third model, firstly, places the words into certain classes by using the co-occurrences of the words, and then uses these classes as the units of the language model. The performance of the models are tested by using two stage decoding; in the first stage, lattices are formed by using bi-gram models and then tri-gram models are used for recognition over these lattices. In this study, it is shown that the vocabulary coverage of the system seriously affects the recognition performance. For this reason, models that use stems and endings as the modeling unit perform better since their coverage of the vocabulary is higher. In addition to that, a single-pass decoder that can perform single pass decoding over these models is believed to increase the recognition performance.

1. Giriş

Günümüzde kullanılan sürekli konuşma tanıma sistemleri çoğunlukla istatistiksel modellemeler kullanılarak yapılandırılır. İstatistiksel konuşma tanıma sistemlerinde akustik ve dil modelleri birlikte kullanılır. Akustik modellemede saklı Markov modelleri kullanılmaktadır. Dil modelleri ise sözcüklerin yan yana bulunma olasılıkları üzerinden oluşturulan N-çekirdekli ("N-gram") modelleri içerir.

Türkçe sondan eklemeli bir dildir. Bu yapısı nedeniyle, zengin biçimbilimsel özelliği bulunmayan dillerde uygulandığı gibi sözcükler üzerinden oluşturulan dil modelleri kullanılamamaktadır. Çünkü, bir sözcüğün her çekimi ayrı bir sözcük olarak düşünüldüğünde, sözcüklerin metin içinde geçme sıklıkları az miktarda olur ve bu şekilde oluşturulan dil modelleri güvenilir olmaz. Ayrıca bu durumda, geniş dağarcıklı tanıma yapabilmek için kullanılması gereken sözcük dağarcığının boyutu çok fazla büyümektedir ve bu kadar büyük sözcük dağarcığı kullanılarak tanıma yapılabilmesi boyutu nedeniyle mümkün değildir.

Bu nedenlerden dolayı geniş dağarcıklı sürekli konuşma tanıma uygulamalarında, eklemeli diller için dil modellenmesinde başka yaklaşımların denenmesi gerekmektedir. Günümüzde diğer eklemeli dillerde dil modellemesi birimi olarak kök, gövde veya ekler gibi sözcük altı birimlerin kullanılmasıyla bu sorunlar aşılmaya calışılmaktadır [1][2][3]. Türkçe için ise sözcük altı birimler ilk defa [4] çalışmasında kullanılmaya başlanmıştır. Başka bir çalışmada ise sözcük altı birimleri ile oluşturulan dil modelleri, biçimbilimsel kuralları içeren ağırlıklı sonlu durum makineleri ile birleştirilmiş ve karma dil modelleri oluşturulmuştur [5].

1-4244-0239-5/06/$20.00

Bu çalışmada ise aynı düşünce doğrultusunda üç çeşit dil modeli incelenmektedir. Bunlardan ilki, sözcüklerin gövdeleri üzerinden oluşturulan dil modelidir. İkincisi ise sözcüğün hem gövdesinin hem de gövdesinden sonra gelen ek dizisinin ayrı birimler olarak alınmasıyla oluşturulmuştur. Son olarak, sözcükler önce sınıflara ayrılmış ve daha sonra bu sınıflar üzerinden dil modeli kurulmuştur. Bu modellerin hepsi kullanılan birimlerin sözcüklerin kendilerinden daha sık bulunması nedeniyle dil modelinin güvenilirliğini artırır. Ayrıca ikinci model, eğitim metninde görulmeyen gövde ek eşleşmelerini de kapsaması nedeniyle sistemin sözcük dağarcığını arttırır.

Yüksek dereceli modelleri sınayabilmek için iki aşamalı tanıma yaklaşımı kullanılmıştır. Bu yaklaşımda önce düşük dereceli bir dil modeli ile tanıma yapılarak en iyi tanınanlar listesi elde edilir ve bu liste üzerinden bir örgü oluşturulur. Daha sonra bu örgü üzerinden yüksek dereceli bir model ile tanıma yapılır. Bu çalışmada "Hidden Markov Model Toolkit" (HTK) [6] kullanıldığından dolayı önce 2-çekirdekli dil modelleri kullanılarak örgüler oluşturulup daha sonra bunlar üzerinden 3-çekirdekli modellerle tanıma yapılmıştır.

2. Veritabanı

Akustik modellerin eğitiminde kullanılan kayıtlar Orta Doğu Teknik Üniversitesi'nde kaydedilmistir. Bu veritabanı 104'ü erkek, 89'u bayan olmak üzere 193 kişinin ses kaydından oluşmaktadır. Her konuşmacı yaklaşık olarak 40 cümle okumuştur, yanlış okumaların bir kısmı veritabanından bütünüyle silinmiştir. Bu işlemin sonucunda veritabanında toplam 6935 cümle kalmıştır [7].

Dil modellerinin oluşturulması için gazetelerin spor sayfalarından çeşitli yazılar toplanmıştır. Oluşan metnin bir kısmının yazım hataları düzeltilmiş, düzeltilmesi mümkün olmayan cümleler ise tamamen silinmiştir. Ayrıca bu metinden rastgele sınama cümleleri seçilmiş ve bu cümleler metinden silinmiştir. Daha sonra bu metindeki sözcükler biçimbilimsel kurallar kullanılarak gövde ve eklerine ayrılmıştır. Bu ayrım sonucunda oluşan birden fazla biçimbilimsel ayrım için en kısa gövdeler seçilmiştir; ama çok sık rastlanan sözcükler için doğru seçim el ile yapılmıştır. Bu işlem sonucunda oluşan metnin istatistiksel bilgileri aşağıda verilmiştir.

Tablo 1: Dil modeli için kullanılan metnin istatistikleri

Sözcük Sayısı	6,507,742
Cümle Sayısı	534,666
Farklı Sözcük Sayısı	161,013
Farklı Gövde Sayısı	45,293
Farklı Ek Dizisi Sayısı	4,622
Farklı Gövde ve Ek Dizisi Sayısı	49,915

Sistemin başarısını sınamak için kullanılan ses kayıtlarını oluşturmak üzere toplanan spor metninden rastgele cümleler seçilmiştir. Bu cümleler dil modeli eğitimi için kullanılmayacağından dolayı spor metininden silinmiştir. Bu cümleler, akustik model eğitiminde kullanılan veritabanında da konuşmacı olarak yer alan 5'i erkek 1'i bayan 6 kişi tarafından okunmuştur. Hatalı okumalar kaydedilen cümlelerden çıkartıldığında geriye 223 cümle kalmıştır. Bu veride yer alan farklı sözcük sayısı 1288'dir.

3. Dil Modelleme

İstatistiksel dil modelleri bir cümlenin dilde bulunma olasılığının kestiriminde kullanılır. Bir cümlenin olasılığı, $\boldsymbol{W}$'nin cümleyi w_i'nin ise o cümlenin bir sözcüğünü gösterdiğini varsayarsak ($\boldsymbol{W} = w_1, w_2, \ldots, w_n$), şöyle yazılabilir [8]:

$$P(\boldsymbol{W}) = \prod_{i=1}^{n} P(w_i | w_1, \cdots, w_{i-1})$$

Kolaylıkla anlaşılabileceği gibi dil modeli bu şekilde oluşturulursa metin içinde cümleler çoğunlukla bir defa geçeceği için sağlıklı bir modelleme yapılamaz. Bu nedenle dil modellerinde sözcük olasıkları hesaplanırken, o sözcükten önce gelen belli sayıda (N) sözcük kullanılır, ve böylece daha sağlıklı bir modelleme yapılmış olur. Bu metoda N-çekirdekli dil modellemesi denir. Daha önce bahsedildiği gibi sürekli konuşma tanıma uygulamasında en sık kullanılan dil modelleri N-çekirdekli dil modelleridir. N-çekirdekli dil modeli ise şu şekilde yazılabilir:

$$P(\boldsymbol{W}) = \prod_{i=1}^{n} P(w_i | w_{i-N+1}, \cdots, w_{i-1})$$

Bu modelleme metodunda "N" 2 seçilirse 2-çekirdekli, 3 seçilirse 3-çekirdekli dil modeli oluşturulmuş olur. Dil modellerinin daha sağlıklı olması için çeşitli yumuşatma metodları uygulanabilir. Bu metodlar, hiç görülmeyen veya az görülen sözcüklerin olasılıklarını yumuşatarak dil modelini geliştirirler.

Bu çalışmada üç çeşit dil modeli incelenmektedir. Bunlardan ilki sözcük gövdeleri üzerinden eğitilen dil modelleridir; bu modeller sözcüklerin sadece gövdelerinin hesaba katılmasıyla oluşturulur. Bu dil modelleri, 2-çekirdekli modelleme için söyle örneklenebilir:

Örnek cümle: *"Ben onların sözünün dışına çıkmam."*

Sözcük gövdeleri: *"ben onlar söz dış çık"*

2-çekirdekli dil modeli olasılıkları: *P(onlar | ben), P(söz | onlar), P(dış | söz), P(çık | dış)*

Gövde modelleri üzerinden tanıma yapılırken eklerin tanınması sadece akustik modeller kullanılarak yapılır. Dolayısıyla bu modeller, bir sözcükten sonra gelebilecek çekim eklerinin dilde o sözcükle görülme olasılıklarını modelleyemez. Bu nedenle de, eklerin tanınması sadece akustik modeller kullanılarak yapılır.

Kullanılan ikinci model gövde ve bu gövdeden sonra gelen ek dizileri üzerinden eğitilir. Bu modeller hem gövde, hem de çekim eklerinin dildeki görülme olasılıklarını modelleyebilir. Bu modeller ise 2-çekirdekli modelleme için

söyle örneklenebilir:

Örnek cümle: *"Ben onların sözünün dışına çıkmam."*

Sözcük gövde ve ek dizileri: *"ben onlar ın söz ünün dış ına çık amam"*

2-çekirdekli dil modeli olasılıkları: *P(onlar | ben), P(ın | onlar), P(söz | ın), P(ünün | söz), P(dış | ünün), P(ına | dış), P(çık | ına), P(amam | çık)*

Bu çalışmada yer alan bir diğer model ise sınıf tabanlı dil modelidir. Bu dil modelinde öncelikle metin içinde geçen sözcükler birbirleriyle görülme olasılıklarına göre istatistiksel olarak sınıflandırılır, daha sonra dil modeli bu sınıflar üzerinden oluşturulur. Bu modellerde, aynı sınıf içinde bulunan sözcüklerin birbirinden ayrılmasında sadece akustik modeller kullanılır. Bu yaklaşım 2-çekirdekli modelleme için söyle örneklenebilir:

Örnek cümle: *"Ben onların sözünün dışına çıkmam."*

Sözcüklerin varsayılan sınıfları: *"S1 S4 S4 S3 S2"*

2-çekirdekli dil modeli olasılıkları: *P(S4 | S1), P(S4 | S4), P(S3 | S4), P(S2 | S3)*

Dil modellemesi sırasında karşılaşılan bir diğer sorun ise sözcük dağarcığının seçimidir. Dil modeli eğitiminde kullanılan metinde geçen tüm sözcükler sözcük dağarcığına konamaz. Özellikle Türkçe'de sözcük sayısı artışı çok hızlıdır [9]. Bu hızlı artış, sözcük dağarcığının kapsamına bir sınır getirir ve sözcük dağarcığı dışında kalan sözcük oranınını ("Out Of Vocabulary rate" — OOV) arttırır.

4. Deneyler ve Sonuçları

Bu çalışmada kullanılan tüm modellerin eğitimi HTK kullanılarak yapılmıştır. Ayrıca tanıma deneyleri yine bu araç kullanılarak yapılmıştır. HTK sürekli konuşma tanıma işlemi sırasında sadece 2-çekirdekli modellerin kullanılmasına olanak vermektedir. Yüksek dereceli modeller kullanılabilmesi için iki aşamalı tanıma yapılması gerekmektedir. İki aşamalı tanıma yapılırken ilk aşamada 2-çekirdekli dil modelleri kullanılarak örgüler oluşturulur ve daha sonra yüksek dereceli dil modelleri kullanılarak bu örgüler üzerinde tekrar tanıma yapılır. Dolayısıyla, basit bir model kullanarak büyük boyuttaki arama uzayı küçültülür ve daha sonra bu daraltılmış uzay üzerinde yüksek dereceli dil modelleri ile tanıma yapılır.

Deneylerin başarıları cümle tanıma oranı, sözcük tanıma oranı ve hassasiyet kriterleri üzerinden ölçülmüştür. Bu kriterlerin tanımı aşağıda verilmiştir.

N = Toplam cümle sayısı

H = Doğru tanınan cümle sayısı

S = Hatalı tanınan sözcük sayısı

D = Sözcüğü yanlışlıkla silme sayısı

I = Araya fazladan koyulan sözcüklerin sayısı

$$\text{Cümle tanıma oranı (SRR)} = \frac{H}{N} \times 100\%$$

$$\text{Sözcük tanıma oranı (WRR)} = \frac{N - S - D}{N} \times 100\%$$

$$\text{Hassasiyet (ACC)} = \frac{N - S - D - I}{N} \times 100\%$$

Önceki bölümde bahsedilen dil modelleri oluşturulmuştur. Sözcüklerin gövdeleri ve sözcük sınıfları üzerinden oluşturulan modeller için sözcük dağarcığı seçimi yapılmıştır. Bu modeller kullanılırken tanıma işlemi sözcükler üzerinden yapılacağından; sözcük dağarcığının, eğitim metnini kapsaması boyutunun büyüklüğü nedeniyle mümkün değildir. Bu nedenle sadece, metinde beşten fazla görülen sözcükler dağarcığa dahil edilmiştir. Bu kısıtlama, bu modeller için kullanılacak sözcük dağarcığının sistem başarısını sınamak için kullanılacak cümlelerdeki sözcüklerin %5,42'sini kapsamamasına neden olmuştur. Gövde ve ek dizileri üzerinden oluşturulan modeller için bu durum geçerli değildir ve bu modeller birkaç sözcük hariç neredeyse sınama cümlelerinde geçen tüm sözcükleri kapsamaktadır. Sözcük kısıtlaması sonucunda sözcük dağarcığına 44.678 tane sözcük konmuştur; gövde ve ek dizileri üzerinden oluşturulan modeller için ise eğitim metninde geçen tüm gövde ve ek dizileri sözcük dağarcığında yer almıştır. Sözcük kısıtlaması açısından başarıyı sınamak için kullanılan cümlelerdeki sözcüklerin dağılımına baktığımızda, cümlelerin %30'unda bir ya da birden fazla dağarcık dışı sözcük bulunduğu görülmektedir. Bu nedenle; gövde ve ek dizileri üzerinden eğitilen modellerin, diğer iki modele göre cümle tanıma kriteri açısından çok daha başarılı olması beklenmektedir.

Konuşma tanıma sisteminin başarısını sınamak için öncelikle 2-çekirdekli dil modelleri üzerinde tanıma işlemi gerçekleştirilmiştir. 2-çekirdekli dil modellerinin başarıları aşağıda verilmiştir.

Tablo 2: 2-çekirdekli dil modellerinin başarı oranları

Model Tipi	*SRR*	*WRR*	*ACC*
Gövde	15.77%	72.38%	67.80%
Gövde ve Ek Dizisi	19.46%	72.36%	70.20%
Sınıf	26.01%	76.53%	74.06%

Yukarıda görüldüğü üzere en başarılı model sınıf tabanlı modeldir. Bu model sözcükleri az sayıda sınıfta toparlayarak, dil modelinin sağlıklı bir şekilde hesaplanmasını sağlamıştır. Gövde ve ek dizilerini birim olarak kabul eden model sadece gövdeleri birim olarak kabul eden modelden daha başarılıdır, bunun nedeni gövde ve ek dizilerini kullanan modelin kapsadığı sözcük dağarcığının daha geniş olmasıdır.

3-çekirdekli dil modellerinin başarısını ölçmek için yukarıda kullanılan herbir 2-çekirdekli dil modeli için en iyi 10 tanıma

sonucu kullanılarak örgüler oluşturulmuştur. Bu oluşturulan örgülerler üzerinde 3-çekirdekli modeller kullanılarak bu modellerin başarıları ölçülmüştür. 3-çekirdekli dil modellerinin başarıları Tablo 3'te verilmektedir.

Tablo 3: 3-çekirdekli dil modellerinin oluşturulan örgülerler üzerindeki başarıları

Model Tipi	*SRR*	*WRR*	*ACC*
Gövde	17.12%	73.46%	69.42%
Gövde ve Ek Dizisi	30.32%	78.54%	76.38%
Sınıf	26.13%	77.78%	74.92%

Tablo 3'te görüldüğü gibi en iyi başarıyı gövde ve ek dizisi üzerinden eğitilen model göstermiştir. Bunun en büyük sebebi bu modelle oluşturulan sözcük dağarcığı kapsamının daha geniş olması ve sınama verisinde geçen sözcüklerin neredeyse tamamını kapsamasıdır. Gövde tabanlı ve sınıf tabanlı modellerde ise ilk aşamaya oranla kayda değer bir artış gözlenememektedir. Buradan da, 3-çekirdekli modellemenin gövdeler ve sınıflar üzerinden eğitilen modeller için 2-çekirdekli modellemeden daha fazla önemli bilgi taşımadığı sonucuna varılabilinir.

5. Değerlendirme

Türkçe eklemeli bir dildir ve bu özelliği nedeniyle sürekli konuşma tanıma uygulamalarında kullanılan dil modellerinin İngilizce gibi dillerde uygulananlardan farklı olması gerekmektedir. Bu çalısmada çeşitli dil modelleri incelenmiştir. Bu modellerin bir kısmı sözcük alt birimlerinin dil modellerinde kullanılmasını, diğerleri ise sözcüklerin belli sınıflar altında gruplanarak bu sınıflar üzerinden modellerin eğitilmesini içermektedir.

Sınıf tabanlı modeller ilk aşamada en başarılı modeller olmasına rağmen; ikinci aşama sonunda başarıları çok artmamaktadır. Bunun nedeni sınıflandırma işleminin 2-çekirdekli olasılıklar kullanılarak yapılmış olmasıdır. Dolayısıyla, 3-çekirdekli dil modelleri için bu sınıflar anlamlı değildir. Sınıflandırmanın daha yüksek dereceli olasılıklara göre ya da sözcüklerin dilbilimsel özellikleri göz önüne alınarak elle yapılması, bu modellerin başarılarını arttırabilir. Yine de bu modellerin sözcük dağarcığı kapsamları düşük olduğu için, bu modeller ulaşılması hedeflenen başarıyı yakalayamayacaklardır.

Deney sonuçlarından da görülebildiği gibi en yüksek başarıyı gövde ve ek dizilerini kullanan dil modelleri sağlamıştır. Bunun nedeni sınama cümlelerinde kullanılan sözcüklerin neredeyse tamamını kapsamasıdır. Bu modeller en başarılı modeller olmasına rağmen ilk aşamada başarıları oldukça düşüktür. 2-çekirdekli modeller, sonrasında ek gelen sözcüklerde gövdeler arasındaki ilişkiyi kapsayamamaktadır; dolayısıyla bu, başarıda bir düşüşe neden olmaktadır.

Sonuç olarak, konuşma tanıma sistemlerinin kapsadığı sözcük dağarcı miktarı, sistemin başarısı açısından çok önemlidir. Bu nedenle, sözcük ve ek dizisi kullanan dil modelleri sözcük dağarcığını kapsamaları bakımından umut vericidir, ama bu modellerde beklenen başarı ilk aşamada 2-çekirdekli dil modelleri kullanma zorunluluğu nedeniyle sağlanamamaktadır. Bu sorunu çözmek için 3-çekirdekli dil modelleriyle tek aşamada tanıma yapabilen sistemlerin hazırlanması gerekmektedir.

6. Kaynakça

[1] Siivola V., Hirsimaki T., Creutz M. and Kurimo M., "Unlimited Vocabulary Speech Recognition Based on Morphs Discovered in an Unsupervised Manner", *Proceedings of the 8th European Conference on Speech Communication and Technology*, 2003.

[2] Siivola V., Kurimo M. and Lagus K., "Large Vocabulary Statistical Language Modeling for Continuous Speech Recognition in Finnish", *Proceedings of the 7th European Conference on Speech Communication and Technology*, 2001.

[3] Byrne W., Hajic J., Ircing P., Jelinek F., Khudanpur S., Krbec P. and Psutka J., "On Large Vocabulary Continuous Speech Recognition of Highly Inflectional Language - Czech", *Proceedings of the 7th European Conference on Speech Communication and Technology*, 2001.

[4] Çiloğlu T., Çömez M. ve Şahin S., "Takılı Bir Dil Olarak Türkçe İçin Dil Modelleme", *12. IEEE Sinyal İşleme ve İletişim Uygulamaları Kurultayı*, 2004.

[5] Büyük O., Erdoğan H., ve Oflazer K., "Konuşma Tanımada Karma Dil Birimleri Kullanımı ve Dil Kısıtlarının Gerçeklenmesi", *13. IEEE Sinyal İşleme ve İletişim Uygulamaları Kurultayı*, 2005.

[6] Young S., Evermann G., Kershaw D., Moore G., Odell J., Ollason D., Valtchev V. and Woodland P., "*The HTK Book (for HTK Version 3.2.1)*", Cambridge University Engineering Department, 2002.

[7] Salor Ö., Pellom B., Çiloğlu T., Hacıoğlu K. and Demirekler M., "On Developing New Text and Audio Corpora and Speech Recognition Tools for the Turkish Language", *Proceedings of 7th International Conference on Spoken Language Processing*, 2002.

[8] Jelinek F., "*Statistical Methods for Speech Recognition*", The MIT Press, 1997.

[9] Çarkı K., Geutner P. and Schultz T., "Turkish LVCSR: Towards better Speech Recognition for Agglutinative Languages" *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2000.