Filler Model Based Confidence Measures for Spoken Dialog Systems
Sesli Diyalog Sistemlerinde Dolgu Modeli Kullanarak Güvenilirlik Ölçümü

Aydın Akyol and Hakan Erdoğan

Sabancı Üniversitesi, Orhanlı Tuzla İstanbul 34956

akyol@su.sabanciuniv.edu, haerdogan@sabanciuniv.edu

Özetçe

Güvenilirlik ölçütü (confidence measure) konuşma tanıma sistemlerinde tanınan sözcüğün gerçekten doğru tanınıp tanınmadığı konusundaki güven düzeyini gösterir. Günümüz ses tanıma sistemlerinin yetersiz performansı nedeniyle, gerçekçi bir güvenilirlik ölçümü, kullanıcı isteklerinin tam olarak anlaşılabilmesi için önemlidir. Bir konuşma tanıma hipotezinin güvenilirliğini belirlemek için elde edilen güvenilirlik öznitelikleri uygun kombinasyonlar halinde bir arada kullanılır. Bu çalışmada, dolgu modeli temelli güvenilirlik öznitelliklerinin performansı incelenmiştir. Bu kapsamda 5 tür dolgu model ağı tanımlanmıştır; üçlü-ses ağı, tek-ses ağı, ses-sınıfları ağı, 5-durumlu genel ses modeli ve 3-durumlu genel ses modeli. Elde edilen özniteliklerin bir Türkçe konuşma tanıma sistemi üzerinde kodçözücü hipotezlerini, tanıma hatası veya doğru tanıma olarak doğru sınıflandırabilme yeteneklerine göre karşılaştırılmışlardır. En iyi başarım, üçlü-ses ağından bulunan öznitelikler ile elde edilmiştir. Ayrıca, özniteliklerin uygun kombinasyonlarının gösterdikleri başarıma bakılmış ve belirli bazı öznitelik kombinasyonlarının güvenilirlik sisteminin başarısını arttırdığı gözlemlenmiştir.

Abstract

Because of the inadequate performance of speech recognition systems, an accurate confidence scoring mechanism should be employed to understand the user requests correctly. To determine a confidence score for a hypothesis, certain confidence features are combined. In this work, the performance of filler-model based confidence features have been investigated. Five types of filler model networks were defined: triphone-network, phone-network, phone-class network, 5-state catch-all model and 3-state catch-all model. First all models were evaluated in a Turkish speech recognition task in terms of their ability to correctly tag (recognitionerror or correct) recognition hypotheses. Here, the best performance was obtained from triphone recognition network. Then the performance of reliable combinations of these models were investigated and it was observed that certain combinations of filler models could significantly improve the accuracy of the confidence annotation.

1. Giriş

Konuşma tanıma kullanılan sesli diyalog sistemlerinde, kullanıcı ile sistem sohbet ederek bir amaca ulaşmaya çalışırlar. Kullanıcıdan gelen isteklerin tam olarak doğru anlaşılması diyaloğun amacına ulaşması için kritik önem taşımaktadır. Öte yandan da günümüzde kullanılan konuşma tanıma sistemlerinin performansı bu seviyede bir doğrululuğu sağlamaktan oldukça uzaktır. Ayrıca konuşma tanıma sistemlerinin %100 doğrulukla çalıştığı durumlarda bile yine de sağlam güvenilirlik ölçütlerine gereksinim duyulacaktır. Çünkü sistemlerin kelime dağarcıklarının dışındaki kelimeleri veya anlam taşımayan sesleri de yakalayıp, doğru söylevlerden ayırt edebilmeleri gerekmektedir. Tipik bir güvenilirlik yaklaşımı iki adımdan oluşur. İlk önce tanıma güvenilirliği ile ilgili olduğu düşünülen bir ya da daha fazla güvenilirlik özniteliği gruplandırılarak bir güvenilirlik vektörü oluşturulur. Daha sonra, bu vektörlere bir ya da daha fazla sınıflandırma tekniği uygulanarak tanımanın güvenilirliği konusunda bir yargı üretilir. Bu açıdan bakıldığında bir güvenilirlik belirleyici sistemin performansını belirleyen en önemli faktörün vektörleri oluşturmada kullanılan ve eğitim verilerinden elde edilen özniteliklerin kalitesi olduğu söylenebilir. Bu yüzden bilgi verici öznitelikler tanımlayarak ve bunların uygun bir kombinasyonunu oluşturarak, güvenilirlik belirleyicisinin performansını artırmak mümkündür. Literatürde güvenilirlik konusunda pek çokpek çok bilgi verici öznitelik tanımlanmıştır [1,2,3,4,5,6]. Bu öznitelikler akustik model, dil modeli, anlambilimsel, tanıma örüsü, ya da en iyi n sonuç listesinden [5] elde edilebilir. Akustik özniteliklerin kullanımı pek çok konuşma sisteminde oldukça başarılı olmuştur. Bu çalışmada biz de akustik özniteliklere ağırlık verdik ve tanımladığımız bazı dolgu modellerine dayanan akustik özniteliklerin, güvenilirlik ölçümü performansına etkilerini inceledik.

En önemli güvenilirlik özniteliğinin, normalize edilmiş kodçözücü skoru olduğu daha önce yapılmış çalışmalar tarafından gösterilmiştir [3,4,7,8,9]. Skorun normalize edilmesi dolgu modelleri ile sağlanabilir. Başka bir deyişle, dolgu modeli tüm akustik alt-birimler için ortak bir Gizli Markov Modeli (GMM) tanımlar. Dolgu modeli bir akustik alt-birim tanıyıcısı olarak görev yapar. Akustik gözlemler sözcük dağarcığında olmasa bile, dolgu model ağı, hipotez ettiği alt-birimleri birleştirerek akustik giriş verisi için uygun bir karşılık bulacaktır. Buna karşın normal kodçözücü en iyi eşleşmeyi sadece kelime modellerine bakarak bulmaya çalışır. Bu noktada şöyle bir sonuç çıkarılabilir; dolgu modelleri yanlış tanımaları ortaya çıkartmak konusunda önemli bilgiler sağlayabilirler.

Bu çalışmada farklı detay seviyelerindeki dolgu model ağlarının güvenilirlik kestirimdeki etkileri incelenmiştir. Farklı dolgu ağlarından elde edilen özniteliklerin birarada kullanılması da incelenmiştir. İncelemelerimizde dolgu modellerden elde edilen ve diğer akustik öznitelikleri kapsayacak en uygun kombinasyonu bulmayı amaçladık.

Makale şu şekilde düzenlenmiştir; 2. kısımda dolgu modeli ağlarının temelindeki teori aktarılarak bir giriş yapılmakta, 3. kısımda ise yaptığımız denemelerle ilgili detaylar sunulmaktadır. 4. kısım sonuçları içermekte ve son kısımda da çalışmanın bir özeti ve gelecek çalışmalar için fikirler ve öneriler yer almaktadır.

0-7803-8318-4104/$20.00©2004 IEEE

2. Güvenilirlik Ölçütü ve Dolgu Modelleri

$O = o_1, o_2, ..., o_t$ akustik gözleminin $W = w_1, w_2, ..., w_n$ kelime dizisi tarafından üretildiğini kabul edelim. Bir konuşma tanıma sisteminin amacı, akustik gözlem sinyali O verildiğinde, en olası kelime dizilimi $\hat{W}$ 'yu tespit etmektir ve bu ifade karşılığını aşağıdaki Bayes' denkleminde bulur;

$$\hat{W} = \arg\max_W P(W|O) = \arg\max_W \frac{P(O|W)P(W)}{P(O)}. \quad (1)$$

Çoğu tanıma sisteminde payda $P(O)$, gözlemin akustik olasılığı, tüm gözlemler için eşit kabul edilir ve hesaplara da katılmaz. Bunun anlamı, tanıma sisteminin ürettiği olabilirlik değerlerinin, $P(W|O)$ mutlak ölçüsünü vermemesidir. Bayes' denkleminin (1) paydasının hipotezlerin güvenilirliklerinin hesaplanmasına katkıda bulunması olasıdır çünkü böylece oran, kelime dizisinin göreli olasılığının $P(W|O)$ mutlak bir ölçüsü olmuş olacaktır.

$P(O)$ 'nun dolgu modellere dayanan genel amaçlı tanıyıcı sistemlerle yaklaşık olarak bulunması mümkündür. Bu tür tanıyıcı sistemler herşeyi tanıyabilmelidir, bu yüzden bir konuşma tanıma sisteminde gramerin kapsamadığı sesleri de doldurabileceklerdir. Bunu yapmak için, öncelikle akustik uzay belirli küçük alt birimlere parçalanıp modellenir ve bu modeller bir kukla gramer yapısı (paralel bağlanmış parçalar ve bir döngü) içerisinde birbirlerine bağlanırlar. Böylece tanıma işlemi kelime dizisi grameri gibi kısıtlanmış ağların getirdiği etkilerden de uzak tutulmuş olur. Başka bir deyişle, dolgu modeller bu birim ağlarından en iyi bağımsız/kısıtlanmamış akustik yolu çıktı olarak verir. $P(O)$'nun değeri aşağıdaki yaklaşıklama ile kestirilebilir:

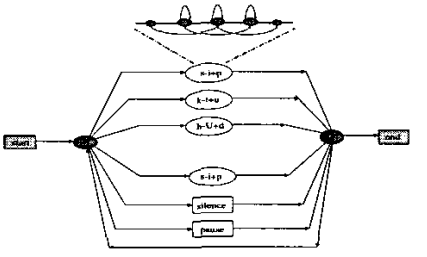
$$\begin{aligned} P(O) &= \sum_U P(O|U)P(U) \\ &\approx \max_U P(O|U)P(U) \\ &\approx \max_{N,u_1 \cdots u_N} \prod_{i=1}^{N} \frac{1}{M} P(O_{t_i}^{t_{i+1}}|u_i). \end{aligned} \quad (2)$$

Burada, U olası tüm birim dizilimlerini göstermekte ve $(u_i)_1^N$ da U'da bulunan N tane birimin listesidir. Ayrıca M dolgu modeli ağındaki birim sayısı yani model sayısı ve $O_{t_i}^{t_{i+1}}$ 'de u_i birim modeli ile hizalanmış gözlem dizisini göstermektedir. Uygulamada, (2)'deki $P(O)$, dolgu modeli ağı üzerinde yapılacak Viterbi kodçözme işlemi ile bulunur. Genelde, bir söz öbeğinin güvenilirliği Bayes' oranı (1) ile ilintili olacaktır. Bu oran normal kodçözünün bulduğu en iyi yolun olabilirlik değeri, yani yaklaşık olarak $P(O|W)P(W)$, ile dolgu modelinin çıkarttığı en iyi yolun olabilirliği, yani $P(O)$, birbirine oranlanarak yaklaşık olarak bulunur. İdeal durumda bu oranın birden küçük olması beklenir fakat pratikte olmayabilir.

Dolgu modeli olarak genelde tek-ses (monophone) ağları ya da genel model (catch-all) ağları kullanılmaktadır ve bu ağlardan gelen çıktı da kodçözücünün bulduğu skoru normalize etmekte kullanılmaktadır. Bizim yaklaşımımızda bunlara ek olarak üçlü-ses (triphone) ağlarının ve ses-sınıfları (phone-class) ağlarının da kullanımını önermekteyiz. Beklenebilir ki, akustik uzayı modellemedeki detay arttırıldıkça, $P(O)$ 'ya daha çok yaklaşılacaktır. Bu noktada, üçlü-ses ağından elde edilecek sonuç, tek-ses ağının üreteceği sonuçtan güvenilirlik belirlemede daha iyi olacaktır. Buna ek olarak, bu tür özellikleri etkin bir sınıfladırıcı kullanarak birleştirdiğimizde performansta önemli artışların olacağı da beklenebilir.

2.1. Üçlü-Ses Tanıma Ağı

Bu tür bir dolgu ağında akustik uzay üçlü-ses'ler kullanılarak modellenir. Üçlü-ses modelleri en önemli eklemsel (coarticulatory) etkenleri yakalayabildikleri için güçlü modellerdir. Türkçe için ses kümemiz 31 sesden oluşmaktadır. Eğitim veri kümemizde yer alan tüm üçlü-ses'leri içeren bir dolgu modeli ağı oluşturmak için öncelikle kelime dağarcığındaki tüm üçlü-ses yapıları belirlendi. Daha sonra eğitim verisine aşırı uyumu (overtraining) azaltabilmek için bir birim kümeleme tekniği uygulandı ve böylece kestirimi yapılacak parametre sayısı 1673 üçlü-ses modeline (her bir üçlü-ses GMM 'si 5-durum[1]/5-karışım topolojisine sahip) ve bir genel duraklama modeli ile bir sessizlik (silence) modeline indirgenmiştir. Şekil 1 bu dolgu modeli gramerini göstermektedir.



Şekil 1: Üçlü-ses dolgu ağı

Her girdi normal tanıyıcının yanısıra bu ağdan da geçirilir. Dolgu modeli skoru o kelimeyle hizalanan tüm çerçevelerden (frame) elde edilen log-olabilirlik değerlerinin toplanması ile hesaplanır.

2.2. Tek-Ses Tanıma Ağı

Bu tür ağda, akustik modellemedeki detay seviyesi azaltılmış ve tüm akustik uzay 31 bağlamdan bağımsız ses modeli ile modellenmiştir. Böylece daha önce 1673 ayrı model ile temsil edilen uzay, bu durumda 31 model ile temsil edilmiştir. Yine benzer şekilde bu modeller kukla-gramer şeklinde biraraya getirilmişlerdir ve böylece üçlü-ses ağında 25905 olan kestirilmesi gereken karışım (mixture) sayısı bu ağda 465 olmuştur.

2.3. Ses-Sınıfı Tanıma Ağı

Burada akustik detay biraz daha azaltılmakta ve tek-ses model ağında kullanılan 31 model, dilbilimsel karakteristik özellikleri gözönünde bulundurularak 6 gruba toplanmıştır [11]. Bu gruplar Tablo 1'de gösterilmiştir. Bu grupların modelleri eğitildikten sonra yine tam bağlantılı bir ağda bir durma ve sessizlik

[1] Sadece 3 yayıcı durum içerir

Tablo 1: Dolgu model olarak kullanılan Türkçe ses sınıfları

Ses Sınıfının Adı	içerilen Sesler
Duraklar	p,t,ç,k,b,d,c,g
Sürtünmeliler	f,s,ş,v,z,j
Genizsel	m,n
Sıvılar	l,r
Kayganlar	y, ğ, h
Kalın sesliler	a, ı, o, u
İnce sesliler	e, i, ö, ü

modeli ile birlikte birbirlerine aynen üçlü ses modelinde olduğu gibi paralel olarak bağlanılmış ve bir geri döngü eklenmiştir.

2.4. Genel Ses Modelleri

Bu dolgu modeli, dolgu modelleri arasındaki en basit ve en yaygın kullanılandır [1,2]. Buradaki ana fikir tüm akustik çeşitliliği genel ve tek bir model ile temsil etmektir. Bu model için bu çalışmada iki farklı topoloji (farklı durum sayıları ve farklı karışım sayıları) denenmiştir. İlkinde 5 durumlu ve diğerinde de 3 durumlu yapı kullanılmıştır.

3. Deneysel Kurulum

Önerilen güvenilirlik ölçütleri Sabancı Üniversitesi Otomatik Ders Sorgulama Sistemi üzerinde denenmiştir. Özetle, sistem telefondan öğrencilerin dersler hakkında sorular sormalarına olanak verir, mesela dersin öğretim üyesini, yerini, saatini, kredisini vs. Bu sistem önceden tanımlanmış bir gramer ve üçlü-ses GMM modelleri kullanarak çalışmaktadır. Normal tanıyıcıyı ve dolgu modellerini eğitmek için iki ayrı veritabanı kullanılmıştır. İlki genel amaçlı bir Türkçe telefon konuşma veritabanı olan TurTel [11] dir. TurTel telefon üzerinden üç farklı mikrofon türü ile toplanmıştır. İçeriği ise Türkçe'nin istatiksel olarak üçlü-ses modellenmesi ile belirlenmiş ve Türkçe'nin %80'ini kapsayacak şekilde 1000 üçlü-sesin kullanılması ile oluşturulmuştur. Bu üçlü-ses'ler 15 cümle ve 373 kelime içerisinde yer almaktadır. Veritabanının konuşmacı kümesini ise farklı yaş, ve bölgelerden 57 erkek ve 36 bayan oluşturmaktadır. Kullanılan diğer veritabanı ise ders sorgulama uygulamamız için topladığımız veritabanıdır. Veritabanındaki tüm kayıtlar derslerle ilgili 4500 sorgu cümlesinden oluşmaktadır. 45 okuyucunun oluşturduğu veritabanındaki verinin %50'si eğitim için TurTel ile birlikte kullanıldı ve geri kalanın %30'u sınıflandırıcı eğitiminde ve %20'si de oluşturulan güvenilirlik ölçme sisteminin test edilmesi için kullanıldı.

Bu çalışmada, güvenilirlik belirleme işlemi iki sınıflı bir sınıflandırma problemi olarak kabul edilmiş ve sınıflar da *doğru tanıma* ve *tanıma hatası* olarak belirlenmiştir. Literatürde güvenilirlik belirleme sistemlerinin değerlendirilmesinde kullanılabilecek pek çok kriter tanımlanmıştır; EER, CER, NCE, NERP, bunlardan bazılarıdır [1,6]. Bu çalışmada EER [Equal Error Rate] değeri kullanılmıştır. EER değeri sınıflandırıcının yanlış kabul (False Accept - FA) oranının yanlış red (False Reject - FR) oranına eşit olduğu çalışma noktasını ifade etmektedir. Bu nokta, Karar Vericinin Etkinlik (Receiver Operating Characteristic - ROC) eğrisinde FA ve FR eksenlerinin orijinine en yakın olan noktadır. ROC eğrisi bir güvenilirlik ölçütünün performansını göstermede kullanılır. Bu eğri FA-FR düzleminde farklı karar/sınıflandırma eşik değerleri için sistemin çalışma değerlerini izler.

Denemelerimizde her bir tanınan kelime için 12 ayrı aday öznitelik çıkarılmıştır. Tüm öznitelikler akustik ve kelime seviyesinde özniteliklerdir. Bunların çoğu (12'nin 8'i) paralel olarak çalışan dolgu modeli kodçözücülerin hipotez edilen kelimenin sınırları kapsamındaki skorlarından çıkartılmıştır.

1. Çerçeve başına log-olabilirlik değeri / Per-frame log-likelihood Score (LL)
2. Üçlü-ses ağı çerçeve başına log-olabilirlik değeri (TL)
3. Tek-ses ağı çerçeve başına log-olabilirlik değeri (PL)
4. Ses-sınıfı ağı çerçeve başına log-olabilirlik değeri (CL)
5. 5-durumlu genel ses modeli çerçeve başına log-olabilirlik değeri (CF)
6. 3-durumlu genel ses modeli çerçeve başına log-olabilirlik değeri (CT)
7. Üçlü-ses log-olabilirlik oran değeri ($TR = LL - TL$)
8. Tek-ses log-olabilirlik oran değeri ($PR = LL - PL$)
9. Maksimum çerçeve skoru (MA)
10. Minimum çerçeve skoru (MI)
11. Hipotezdeki çerçeve skorlarının standart sapması (SD)
12. Hipotez edilen kelimedeki ses sayısı (NP)

Burada, çerçeve seviyesinde değerler türettiğimiz için tüm değer hesaplamalarında normalize edilmiş birim değerlerini kullandık ve bu yüzden tüm özellikler *çerçeve başına* değer olarak adlandırılmıştır.

4. Sonuçlar

İlk önce belirli dolgu model ağ özniteliklerinin bireysel olarak performansları incelenmiş ve böylece seçilen dolgu modeli ağlarının kalitesi konusunda fikir verilmeye çalışılmıştır. Bu nedenle 5 temsili öznitelik seçilmiş ve EER değerleri hesaplanmıştır. Sonuçlar Tablo 2'de gösterilmiştir. Basit bir tek boyutlu Gauss Karışım Modeli (GKM) sınıflandırıcısı kullanılmış ve böylece modeller her bir sınıfta bir Gauss karışımı ile temsil edilmiştir. Gauss karışımları, sınıflandırıcı eğitimi için kullanılmak üzere ayrılmış olan veritabanı ile eğitilmiştir. Bu amaçla, eğitim verisindeki her kelime *doğru* veya *yanlış* olarak etiketlenmiştir. Test işlemi için, GKM'lerden elde edilen hipotezlerin olabilirlik oranları (likelihood ratio) hesaplanmış ve değişen karar eşik değerleri ile karşılaştırılmıştır (likelihood ratio test). Her bir EER değer hesaplanmasında 40 farklı karışım kombinasyonu denenmiş ve her bir kombinasyonada 140 farklı eşik değeri uygulanmıştır.

Tablo 2: Bireysel dolgu model ağ perfromans sonuçları

Dolgu Model Tipi	EER(%)
Üçlü-ses log-olabilirlik oran değeri - (TR)	27.12
Tek-ses log-olabilirlik oran değeri - (PR)	33.49
Ses-sınıfı tanıma ağı - (LL,CL)	38.78
5-durumulu genel ses modeli - (LL, CF)	38.37
3-durumulu genel ses modeli - (LL, CT)	42.78

Tablo 2'deki sonuçlara göre detaylı dolgu ağlarındaki akustik modellemenin detayı arttırıldıkça daha iyi sonuçlara, yani daha düşük EER değerlerine ulaşılabileceği yargısına varılabilir. Ayrıca bireysel performanstaki azalmaya karşın her bir dolgu model tipi, tanımanın güvenilirliği hakkında farklı bilgiler katabilir. Başka bir deyişle, tüm dolgu model değerleri örtüşen bilgiler içerebilirler ama öte yandan da diğer model tiplerinden elde edilemeyecek bazı özel güvenilirlik bilgilerini de içeriyor olabilirler. Örtüşen ve ayırdedici güvenilirlik bilgileri arasındaki ikilemi anlayabilmek amacıyla Tablo 3'de farklı öznitelik kombinasyonlarının performans değerleri sunulmuştur (Aslında 60 farklı kombinasyon incelenmiştir ama burada sadece önemli 14 tanesinin sonuçları verilmiştir).

Tablo 3: GKM sınıflandırıcı için öznitelik kombinasyon sonuçları.

Öznitelikler	EER(%)	Öznitelikler	EER(%)
TR	27.82	LL,TR,PR	26.52
PR	33.49	LL,MA,MI	37.03
LL,CL	38.78	TR,PR,NP	24.32
LL,CT	42.78	CL,TR,PR	25.35
LL,TR	26.98	LL,TR,PR,SD	25.55
LL,PR	34.42	CL,CF,TR,PR	26.23
TR,PR	25.62	CL,CF,TR,PR,SD,NP	26.73

Bireysel performanslar arasında öznitelik TR, *üç-sesli log-olabilirlik oranı*, en iyi EER değerini verir, %27.82. Tablodaki en iyi sonuç olan Kombinasyon (TR,PR,NP): %24.32 ile arasındaki fark sadece %3.50 'dir ve ayrıca bu öznitelik bu kombinasyon tarafından da içerilmektedir. Tablodaki en iyi sonucu veren kombinasyondaki bilginin büyük bölümü TR tarafından sağlanmaktadır. Öznitelik TR 'nin performans sonucunun iyi çıkması aslında beklenilen bir sonuçtu, çünkü öznitelik TR kısıtlanmamış konuşmanın en detaylı modellenmesinin bilgisini kullanır.

Tablodaki en iyi EER değeri olan %24.32'ye sahip kombinasyon (TR,PR,NP), TR'ye göre %3.5 gibi önemli bir performans artışı sağlamaktadır. Eğer (TR,PR,NP) kombinasyonundaki TR özniteliği dışındaki diğer öznitelikleri çıkartmanın maliyetinin TR'yi bulmanın maliyetinin çok çok altında olduğunu düşünürsek sağlanan bu iyileştirmenin önemi daha da iyi anlaşılacaktır. Öznitelik TR'nin çıkartılması için 1675 modelden oluşan bir ağda arama yapılırken örneğin öznitelik PR için sadece 31 modellik bir ağda arama yapılmaktadır.

En iyi kombinasyon olan (TR,PR,NP)'i içermesine karşın, kombinasyon (CL,CF,TR,PR,SD,NP)'nın performansı %26.73 olarak bulunmuş yani kendi altkümesinden daha kötü sonüç vermiştir. Bunun nedeni bu kombinasyonun sınıflandırmada ayırt ediciliğe sahip olmayan hatta gürültü yaratan ve öznitelik vektörünün toplam kullanılabilirliğini kötü etkileyen öznitelikler içermesidir. Bu olay için başka bir açıklama da aşırı boyutlandırma (curse of dimensionality) problemi olarak gösterilebilir. Yani sınıflandırıcı eğitmekte kullanılan veriler bu 6 boyutlu uzayda etkin bir sınır bulmada yetersiz kalmış olabilir.

5. Özet ve Öneriler

Bu çalışmada sürekli konuşma tanıma sistemleri için farklı detay seviyelerindeki dolgu modellerin güvenilirlik belirlemedeki performansları incelenmiştir.

Bireysel öznitelik performansları açısından incelenen dolgu model tipleri arasında üçlü-ses tanıma ağının literatürde yaygın olarak kullanılan tek-ses model ağına oranla %5.67 gibi avantaj sağladığı görülmüştür. Bu iyileşmedeki temel neden üçlü-ses ağının diğer ağlardakinden daha detaylı modeller kullanmasıdır. Bu detaylı model kullanımı, beraberinde eğitim verisi ve hesaplamalarda artış getirmesine rağmen bu tür dezavantajlar gerçekleştirme sırasında uygulanabilecek bazı küçük hileler ile hafifletilebilir, örneğin budama veya parametre bağlama gibi.

Genelde bir hipotezin doğruluğunun tespitinde tek bir uygun güvenilirlik bilgisi kullanılır. Ama bu çalışmada görülmüştür ki bu tür özellikleri uygun kombinasyonlarla biraraya getirerek güvenilirliği belirlemede %3.5'a varan performas arttırımları elde edilebilmektedir.

Gelecekte, güvenilirlik belirlemedeki performansı artırmaya yönelik hece ağı ya da kukla gramerli tek-sözcük (unigram) ağı gibi alternatif dolgu model tipleri kullanılabilir. Ayrıca dolgu model öznitelikleri ile bu çalışmada bahsi geçmeyen diğer akustik özniteliklerin uygun kombinasyonlarının performansları incelenebilir. Son olarak, Türkçe konuşma tanıma uygulamalarındaki doğruluk oranını artırmak için daha detaylandırılmış bir ses-kümesi kullanmak faydalı olabilir. Oflazer [10] standart 29 ses yerine 45 sesden oluşan yeni bir Türkçe ses-kümesi tanımlamıştır.

6. Kaynakça

[1] Chase L. Error-Responsive Feedback Mechanisms for Speech Recognition. Phd Thesis, CMU, 1997.

[2] Hazen T.J., Burianek T., Polifroni J., Seneff S. Recognition confidence scoring for use in speech systems. CSL16, 49-67, 2002.

[3] Cox S., Dasmahapatra S. A high-level approach to confidence estimation in speech recognition. Eurospeech 1999.

[4] Kamppari S., Hazen T., 2000. Word and phone level acoustic confidence scoring. ICASSP 2000.

[5] Zhang R., Rudnicky A.I., Word level confidence annotation using combinations of features.Eurospeech 2001.

[6] Schaaf T., Kemp T. Confidence measures for spontaneous speech recognition. ICASSP 1997.

[7] Gunawardana A., Hon H. W., Jiang L. Word-based acoustic confidence measures for LVSR. ICASSP 1998.

[8] Weintraub, M. LVCSR log-likelihood ratio scoring for keyword spotting. ICASSP 1995.

[9] Cox S., Rose R. Confidence measures for the switchboard database. ICASSP 1996.

[10] Oflazer K., Inkelas S. A Finite State Pronunciation Lexicon for Turkish. EACL Workshop on FSMs in NLP 2003.

[11] Yapanel U. Garbage modelling techniques for a Turkish keyword spotting system, Msc. Thesis, Boğaziçi University, 2000.