Türkçe Ağız Tanıma İçin Bir Veriseti ve Derin Öğrenme ile Sınıflandırma

A Dataset For Turkish Dialect Recognition and Classification with Deep Learning

Gültekin IŞIK
Bilgisayar Mühendisliği Bölümü
Iğdır Üniversitesi
Iğdır, Türkiye
gultekin.isik@igdir.edu.tr

Harun ARTUNER
Bilgisayar Mühendisliği Bölümü
Hacettepe Üniversitesi
Ankara, Türkiye
artuner@cs.hacettepe.edu.tr

***Özetçe*— Ağız Tanıma Sistemleri (ATS) ağız bölgelerinde bulunan benzer akustik özelliklere göre ağızların gruplandırılmasını sağlayan sistemlerdir. Konuşmacının yaşı, cinsiyeti ve ağız özellikleri konuşma tanıma sistemlerinin performansını olumsuz etkilemektedir. Ağız farklılıklarını işlemek için, ağız tanıma sistemleri konuşma tanıma sistemlerine entegre edilebilir. Konuşulan ağzın belirlenmesiyle sistem, ilgili konuşma tanıma modeline anahtarlanabilir. Türkçede otomatik ağız tanıma sistemlerine yönelik kullanılabilecek bir veri seti yoktur. Bu çalışmada bu eksikliğin bir nebze de olsa giderilmesi düşünülmektedir. Ayrıca, oluşturulan veri setinin konvolüsyonel sinir ağlarıyla sınıflandırılmasına yönelik bir deneysel çalışma yapılmıştır. Sonuçta elde edilen % 83,3 doğruluk oranı tatmin edicidir.**

Anahtar Kelimeler — türkçe ağız tanıma, türkçe ağız veriseti, konvolüsyonel sinir ağları.

***Abstract*— Dialect Recognition Systems (DRS) are systems that group dialects, according to similar acoustic features found in dialect regions. The speaker's age, gender, and dialect characteristics negatively affect the performance of speech recognition systems. To handle dialect differences, dialect recognition systems can be integrated into speech recognition systems. By determining the spoken dialect, the system can be switched to the corresponding speech recognition model. There is no dataset that can be used for Turkish automatic dialect recognition systems. In this study, it is thought that this deficiency should be eliminated in some way. In addition, an experimental study has been carried out to classify the generated data set by convolutional neural networks. The resulting 83.3% accuracy is satisfactory.**

Keywords — turkish dialect recognition, turkish dialect dataset, convolutional neural networks.

I. GİRİŞ

Otomatik Konuşmacı Tanıma (OKT) sistemleri genel itibariyle konuşmacı kimlik saptama (KKS), konuşmacı kimlik doğrulama (KKD) ve otomatik konuşmacı sınıflandırma (OKS) olarak üç başlık altında incelenebilir. KKS sistemleriyle, kayıtlı olan seslerden konuşmacının kimliği belirlenir. KKD sistemi, konuşan kişinin gerçekten iddia edilen kişi olup olmadığını tespit etmek için kullanılır. OKS sistemleri ise kişilerin, ağız bölgelerinde bulunan benzer akustik özelliklere göre gruplandırılmasını sağlar ve buna genelde ağız tanıma sistemleri (ATS) de denilmektedir.

Ağız tanıma sistemleri tamamiyle veriye bağımlıdır. Ses kaydının ortam koşulları, seçilen kişilerin özellikleri ve kullanılan cihazlar ATS'nin performansını doğrudan etkiler. Bu etkenlerin bilinmediği ve standart olmadığı durumlarda bu sistemlerin başarım oranlarının ve diğer sistemlerle karşılaştırılmasının bir anlamı yoktur [1]. Bu çalışmanın amacı Türkçenin dört ağız bölgesinden (Ankara, Alanya, Kıbrıs, Trabzon) toplanan ses kayıtlarının, konuşmacı ve ağız tanıma sistemleri için derlenmesi ve işlenebilir hale getirilmesidir. Bu çalışmada ağız tanıma sistemleri için, metne dayalı olmayan (spontane) bir konuşma veriseti oluşturulmasına yönelik yöntem sunulmuş ve bu veriseti kullanılarak bir deneysel çalışma gerçekleştirilmiştir.

Türkçenin ağızlarına yönelik, bu çalışmada bahsedilen şekliyle bir veriseti yoktur ve olanlar da daha çok ağız bilimi çalışmalarıyla sınırlı olacak şekildedir. Bu çalışmada ağız bilimi çalışmalarına da yardımcı olması beklenilen bir Türkçe veriseti tanıtılmaktadır. Ancak Avrupa dillerine ve özellikle İngilizce diline yönelik yapılmış çalışmalar mevcuttur. TIMIT veriseti [2] Amerika'da bulunan 8 ağız bölgesinden ses kayıtlarını ve bunların fonetik düzeyde etiketlerini içeren bir verisetidir. Amerikan [3], Mısır [4] ve Hint [5] ağızlarına yönelik yapılmış CallFriend, yine aynı şekilde Amerikan [6], Mısır [7] ve İnpanyol [8] ağızları için yapılmış CallHome verisetleri vardır. Bunların yanında İskandinav ağızlarına [9] yönelik olarak oluşturulan çalışmalar bulunmaktadır.

Ağız, aynı kökten geldiği bir standart dilden belli oranda ayrılabilen yerel konuşma biçimi olarak tanımlanmaktadır [10]. Ağızlar, ait olduğu dilden sessel (fonolojik), şekilsel (morfolojik), söz varlığı (leksikal) ve söz dizimi (sentaks)

978-1-5386-1501-0/18/$31.00 ©2018 IEEE

özellikleriyle ayrılır [11]. Bu yüzden ağız tanıma dil tanımanın özel bir durumu olarak düşünülebilir. Konuşma tanıma sistemlerinin performansını etkileyen en önemli faktörler arasında konuşmacının yaşı, cinsiyeti ve ağız farklılıkları sayılabilir [12]. Bundan dolayı konuşmacıdan bağımsız konuşma tanıma sistemlerinde ağız farklılıklarının ele alınması gerekir [13]. Böylece konuşma tanıma sistemi, konuşmacının ağız bilgisine göre ilgili modele anahtarlanabilir.

Derin Öğrenme, derin mimariye dayalı yapay sinir ağları ve bunlar için geliştirilen algoritmalar olarak tarif edilebilir [14]. Bu mimaride, her katmanda sınıflanması beklenen verilere ait öznitelikler öğrenilir ve öğrenilen bu öznitelikler bir sonraki katmanda kullanılır. Böylece giriş katmanından çıkış katmanına doğru en basitten en karmaşık öznitelikler öğrenilmiş olur.

Konvolüsyonel Sinir Ağları (Convolutional Neural Networks, CNN) konuşma [15] ve görüntü [16] sınıflandırmada başarıyla uygulanmaktadır. Bu çalışmada ses kayıtlarının mel spektrogramları çıkartılmış ve CNN mimarisinde kullanılarak sınıflandırma yapılmıştır.

Çalışmanın ikinci bölümünde veri setiyle ilgili oluşturulma süreci ve sayısal bilgiler verilmiş, üçüncü bölümde kullanılan metotlar tanıtılmış, dördüncü bölümde yapılan uygulamaya yer verilmiştir. Beşinci bölüm uygulamaya ait sonuçların tartışılmasına ve gelecekte yapılması planlanan işlere ayrılmıştır.

II. TÜRKÇE AĞIZLAR VERİ SETİ

A. *Nasıl Oluşturuldu?*

Veri setinin cinsiyete bağımlı olmaması için eşit sayıda kadın ve erkekten kayıtlar toplanmalıdır. Konuşmacılardan genelde istenen, bir hatırasını anlatması veya gelenek-göreneklerinden bahsetmesidir. Bunlar ne kadar doğal ve uzun konuşmalara dayalı olursa istenilen şartları o kadar sağlamış olur. Uzun konuşmalar, fonetik olarak farklı seslerin elde edilmesi ihtimalini arttırır.

Bu çalışmadaki kayıtlar, genelde mülakat yapılan kişinin evinde yapıldığı için gürültülüdür. Ses özelliklerinin elde edilmesini kolaylaştırmak için kayıtlar bu gürültülerden arındırılmalıdır.

Türkçenin ağız özelliklerinin incelenmesi ve birbiri arasındaki etkileşim ve farkların ortaya konulması için dil bilimciler ilgili ağız yöresine geziler düzenlemektedir. Bu gezilerde, ağız özelliklerini yansıtacağına inanılan kişilerle mülakat yapılır. Bu mülakatlar esnasında, konuşmalar ses kayıt cihazlarıyla kaydedilir. Kaydın yapıldığını ve bunların daha sonra hangi amaçlarla kullanılacağı, mülakat yapılan kişiler tarafından bilinmektedir.

Bu çalışmada tanıtılan veriseti dil bilimcilerden elde edilen içeriklerin düzenlenmesiyle oluşturulmuştur. Türkçenin Ankara, Kıbrıs, Trabzon ve Alanya ağız bölgelerinden toplanan örnekler kullanılmış ve bunlar üzerinde çalışmalar yürütülmüştür. Bu ağızlara sahip yörelerde, ağzın özelliklerini taşıdığı düşünülen kişiler seçilmiştir. Seçilen kişilerin yaşlı olmasına, çoğunlukla bulunduğu yöreyi terk etmemiş olmasına ve eğitim seviyesinin düşük olmasına dikkat edilmiştir. Sayılan bu özellikleri sağlayan kişilerde ağza özgü seslerin bulunma ihtimalinin yüksek olduğu bilinmektedir [17].

Şekil 1. Uygulanan işlemler

B. *Veri Setine İlişkin Sayısal Bilgiler*

Veri seti; gürültü, kanal sayısı farklılığı, örnekleme frekansı farklılığı gibi etkilerden ve sessizlik (silence) bölgelerinden arındırılmıştır. Böylece Ankara için 0,8h, Kıbrıs 0,65h, Trabzon 0,55h ve Alanya 0,7h olmak üzere toplamda dört ağız bölgesinden 2,7h veri elde edilmiştir. Her bir ağız için ikisi kadın ikisi erkek dört kişiden konuşmalar alınmıştır. Tüm kayıtların örnekleme frekansı SoX [18] yazılımıyla 16 KHz'e dönüştürülmüştür. Konuşma kayıtları sözcük seviyesinde Praat [19] yazılımıyla etiketlenmiş daha sonra ortalama 2-3s uzunluklu cümlelere denk gelecek şekilde ayrılmıştır. Sesler ve bunların transkript metinleri ayrı dosyalar halinde kaydedilmiştir. Böylece her ağız bölgesi için yaklaşık 400 cümle belirlenmiştir (Şekil 1). Bu veriseti herhangi bir metne dayanmamakta ve kendiliğinden gelişen konuşmalardan oluşmaktadır. Verisetine ilişkin sayısal bilgiler Tablo 1'de verilmiştir.

TABLO I. VERISETI ÖZET BILGILERI

	Toplam	Ankara	Alanya	Kıbrıs	Trabzon
Konuşmacı	16	4	4	4	4
Sesli ifade	1595	420	410	385	380
Sözcük	3620	935	909	891	885
Uzunluk	2.7h	0.8h	0.7h	0.65h	0.55h

III. KONVOLÜSYONEL SİNİR AĞLARI (CNN)

CNN mimarisi iki aşamalı çalışır. Birinci aşamada birbiriyle ilişkili olan yerel özniteliklerin çıkartılmasını, ikinci aşamada ise çok katmanlı sinir ağları kullanılarak sınıflandırma yapılmasını sağlar. Küçük boyutlu filtrelerin, ağın girdi verisi üzerinde konvolüsyon işlemine tabi tutulmasıyla yerel öznitelik haritaları elde edilir. Bir filtre tüm girdi verisine uygulandığı için ağın bu özelliğine ağırlıkların paylaşılması özelliği denir. Filtrelerin parametreleri, eğitim boyunca hatanın geriye yayılımı algoritmasıyla güncellenir, böylece filtrelerin içerikleri öğrenilmiş olur. Öznitelik haritaları, katmanın çıkışına aktivasyon fonksiyonları uygulanarak elde edilir. Öznitelik haritaları daha sonra *pooling* işlemine tabi tutulur. *Pooling* işlemi öznitelik haritalarının boyutunu düşürür ve böylece özniteliklerdeki çeşitliliği azaltır. *Pooling* işlemi genelde en büyüğü bulma (max) operatörü ile yapılır. Seçilen ebatta *pooling* penceresinin içerisinde kalan özniteliklerin en büyük elemanı bulunur ve öznitelik haritaları tekrar oluşturulur.

CNN mimarisinin sınıflandırma aşaması çok katmanlı sinir ağlarından oluşur. *Pooling* işleminden elde edilen öznitelik haritaları vektör haline getirilerek sinir ağının girişine verilir. Genelde gizli katmanlarda sigmoid aktivasyon fonksiyonu, çıkış katmanında ise softmax fonksiyonu kullanılır. Çıkış katmanında her bir sınıf için sonsal olasılıklar hesaplanarak en yüksek olasılıklı olana karar verilir.

IV. DENEYSEL ÇALIŞMA

Akustik açıdan ağız tanıma, ağızların spektral dağılımları bakımından farklılık göstermesi esasına dayanır. Konuşma sinyalinin fiziksel düzeydeki özellikleri Türkçenin ağızlarını birbirinden ayırt edebilir. Akustik seviyedeki bilgi, sinyalin ham halinden mel spektrogram gibi öznitelikleri çıkartılarak elde edilebilir. Akustik özelliklere dayalı olarak ağızların sınıflandırılması için CNN mimarisi kullanılmıştır.

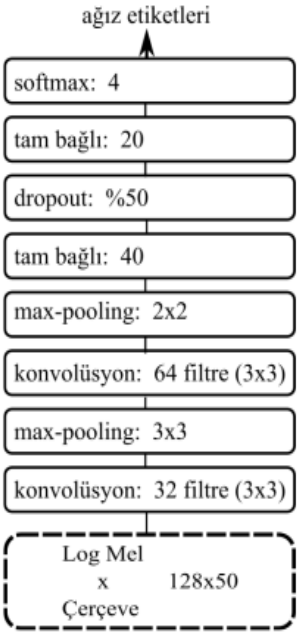
Spektrogram, sinyalin frekans spektrumunun zamana bağlı değişimini ifade eder. Sinyalin frekansı mel ölçeğine çevrilerek logaritmik güç spektrumu elde edilir. Log mel-spektrogram bu güç spektrumunun zamana bağlı değişimini gösterir.

K-katlamalı çapraz doğrulama (k-fold cross validation) yöntemiyle eğitim ve test verisi 10 parçaya ayrıldı ve 10 defa arka arkaya deneme yapıldı. Her defasında 9 parça eğitim, 1 parça da test verisi olarak kullanıldı. Daha sonra bu 10 denemenin ortalaması alınarak sonuç skoru elde edildi.

Açık kaynaklı librosa kütüphanesi [20] kullanılarak her çerçeve için 128 mel özniteliği çıkartıldı. 50 çerçevelik parçalar halinde mel spektrogramlar elde edildi (128x50 ebatlı) ve CNN girişine verildi. CNN mimarisinin parametreleri Şekil 2'de gösterilmektedir.

TABLO II. KARIŞIKLIK MATRISI

Veri seti	Ankara	Alanya	Kıbrıs	Trabzon
Ankara	82,9	5,5	5,4	6,2
Alanya	5,5	84,0	6,0	4,5
Kıbrıs	5,6	5,9	83,4	5,1
Trabzon	6,4	5,5	5,1	83,0



Şekil 2. CNN parametreleri

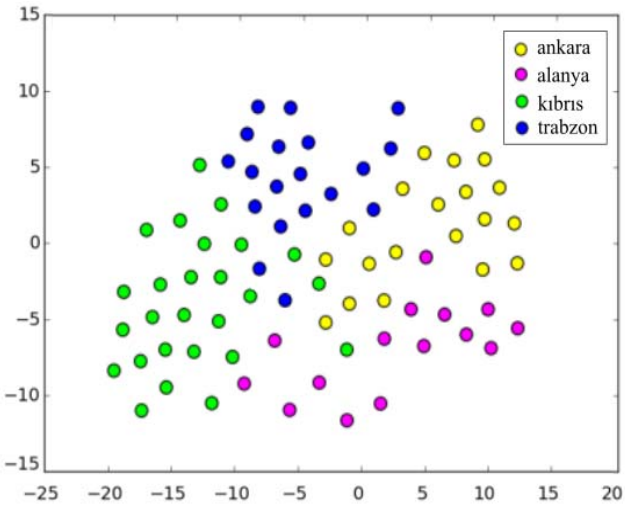
Modelde sırasıyla 32 ve 64 filtreden oluşan iki konvolüsyon katmanı kullanıldı. Bu katmanlardan sırasıyla 32 ve 64 adet öznitelik haritası elde edildi. Mel-spektrogram matrisine 3×3 boyunda filtreler uygulanarak her iki eksen (frekans ve zaman) boyunca da konvolüsyon işlemi yapıldı. Ağda aktivasyon fonksiyonu olarak sigmoid, çıkış katmanında softmax fonksiyonu kullanıldı. *Cross-entropy* hata fonksiyonuna göre eğitim yapıldı. CNN mimarisinin kurulumu ve eğitilmesi için Keras kütüphanesi [21] kullanıldı.

V. SONUÇLAR VE TARTIŞMA

Sınıflandırma sonucunda % 83,3 doğruluk oranı elde edilmiştir. Bu oran, dört sınıf için rastgele sınıflandırma oranından (% 25) çok yüksektir. Tablo 2'de Türkçenin ağızları verisetinin karışıklık matrisi görülmektedir. Bu tabloda satırlar gerçek sınıfı, sütunlar ise modelin ürettiği sınıfı göstermektedir. Buna bakarak Alanya ile Kıbrıs ağızlarının birbirine yakın, Kıbrıs ile Trabzon ağızlarının ise birbirinden uzakta oldukları görülebilir. Değerler yüksekse birbirine karıştırılma oranı yüksektir ve bu da iki ağzın birbirine yakınlığını gösterir. İki ağzın birbirinden tam olarak ayrışması, coğrafik olarak aralarındaki mesafenin uzaklığına da bağlanmaktadır.

Ayrıca Şekil 3'te t-SNE [22] yöntemiyle Şekil 2'deki modelin en üstündeki tam bağlı katmanda elde edilen öznitelikler iki boyutlu olarak gösterilmiştir. Bu şekilde, kullanılan sinir ağının öğrendiği özniteliklerin, Türkçenin ağızlarını ayırt etmedeki başarısı net olarak görülmektedir.

Bu verisetinin toplanmasında bahsedilen yöntemle konuşma verisi elde etmek zahmetli ve maliyetli bir iştir. Bunun yerine, çeşitli sosyal medya araçları, televizyon ve radyo yayınları veya telefon operatörlerinden bu veriler toplanabilir. Her ne kadar telefon operatörlerinin, ellerindeki verileri paylaşması çok zor olsa da diğer yöntemlerle veri toplanmasında bir sınır yoktur.



Şekil 3. CNN'deki son tam bağlı katmanın t-SNE yöntemiyle boyutu düşürülerek görselleştirilmesi

Uygulama arayüzleri (API) kullanılarak özellikle youtube kayıtlarının elde edilmesi mümkündür. Bunun yanında, internet üzerinden akış sağlayan yerel televizyon ve radyo kanallarının içerikleri çeşitli yazılımlarla kaydedilebilir. Böylece, toplanan verilerin boyutu arttırılarak daha büyük uygulamalarda kullanılabilir.

Bunların yanında Türkçenin ağızlarına ilişkin yapılan gezi çalışmalarıyla toplanan veriler azımsanmayacak büyüklüktedir. Her yörenin ağız özelliklerini yansıtan kayıtlar, birbirinden bağımsız ve habersiz ekipler tarafından toplanmaktadır. Bu yüzden, Türkçenin bütün ağızlarını kapsayan otomatik sınıflandırma ve tanıma çalışmaları yapmak zordur. Bu zorluğun üstesinden gelmek için bu çalışmada tanıtılan verisetinin elektronik ortamda herkese açık hale getirilmesi ve isteyenlerin bu verisetine katkısının sağlanması düşünülmektedir. Böylelikle Türkçenin bütün ağızları ve hatta lehçelerine yönelik bir verisetinin oluşturulması planlanmaktadır.

TEŞEKKÜR

Bu çalışmada Türkçe Ağızları Veri Setinin oluşturulması aşamasında, işlenmemiş kayıtlarını bizimle paylaşan, deneyimleriyle yol gösteren Prof. Dr. Nurettin Demir'e teşekkür ederiz.

KAYNAKLAR

[1] Reynolds, D. A., "An overview of automatic speaker recognition technology," in *International conference on acoustics, speech and signal processing, ICASSP'02*, 2002, pp. 4072–4075.

[2] Garofolo, J. S., Lamel, L. F., Fischer, W. M., Fiscus, J. G., Pallett, D. S. and Dahlgren, N. L., "Acoustic-Phonetic Continuous Speech Corpus," vol. 0, no. January 1993, pp. 1–94, 1993.

[3] Canavan, A. and Zipperlen, G., "CallFriend american english," in *Linguistic Data Consortium*, 1996.

[4] Canavan, A. and Zipperlen, G., "CallFriend Egyptian Arabic Speech," in *Linguistic Data Consortium*, 1996.

[5] Canavan, A. and Zipperlen, G., "CallFriend hindi speech corpus," in *Linguistic Data Consortium*, 1996.

[6] Canavan, A., Graff, D. and Zipperlen, G., "CallHome american english speech," in *Linguistic Data Consortium*, 1997.

[7] Canavan, A., Zipperlen, G. and Graff, D., "CallHome Egyptian Arabic Speech," in *Linguistic Data Consortium*, 1997.

[8] Canavan, A. and Zipperlen, G., "CallHome Spanish Speech," in *Linguistic Data Consortium*, 1996.

[9] Johannessen, J. B., "The Nordic Dialect Corpus - an Advanced Research Tool," in *Proceedings of the 17th Nordic Conference of Computational Linguistics*, 2009.

[10] Zissman, M. A., "Comparison of four approaches to automatic language identification of telephone speech," *IEEE Trans. Speech Audio Process.*, vol. 4, no. 1, pp. 31–44, 1996.

[11] Etman, A. and Louis, A. A., "American dialect identification using phonotactic and prosodic features," *IntelliSys 2015 - Proc. 2015 SAI Intell. Syst. Conf.*, pp. 963–970, 2015.

[12] Huang, R., Hansen, J. H. L. and Angkititrakul, P., "Dialect/Accent Classification Using Unrestricted Audio," *IEEE Trans. Audio. Speech. Lang. Processing*, vol. 15, no. 2, 2007.

[13] Biadsy, F., "Automatic Dialect and Accent Recognition and its Application to Speech Recognition," *PhD Thesis,Columbia Univ.*, pp. 1–171, 2011.

[14] Sainath, T. N., Weiss, R. J., Senior, A. W., Wilson, K. W. and Vinyals, O., "Learning the speech front-end with raw waveform CLDNNs," *Interspeech*, 2015.

[15] Abdel-Hamid, O., Mohamed, A.-R., Jiang, H., Deng, L., Penn, G. and Yu, D., "Convolutional Neural Networks for Speech Recognition," *IEEE/ACM Trans. AUDIO, SPEECH, Lang. Process.*, vol. 22, no. 10, 2014.

[16] Cireşan, D. C., Meier, U., Masci, J., Gambardella, L. M. and Schmidhuber, J., "Flexible, High Performance Convolutional Neural Networks for Image Classification," *Proc. Twenty-Second Int. Jt. Conf. Artif. Intell.*, 2011.

[17] Demir, N., "Ağız Araştırmalarında Kaynak Kişi Meselesi," *Folk. Prof. Dr. Dursun Yıldırım Armağanı*, p. 11, 1998.

[18] "Sound eXchange software." [Online]. Available: http://sox.sourceforge.net/. [Accessed: 14-Feb-2018].

[19] Boersma, P. and Weenink, D., "Praat: doing phonetics by computer [Computer program]," 2018. [Online]. Available: http://www.praat.org/. [Accessed: 03-Feb-2018].

[20] Mcfee, B., Raffel, C., Liang, D., Ellis, D. P. W., Mcvicar, M., Battenberg, E. and Nieto, O., "librosa: Audio and Music Signal Analysis in Python," *Proc. 14th Python Sci. Conf.*, no. Scipy, pp. 1–7, 2015.

[21] Chollet, F., "Keras," *Github*, 2015. [Online]. Available: https://github.com/fchollet/keras. [Accessed: 15-Nov-2017].

[22] Van der Maaten, L. and Hinton, G., "Visualizing data using t-sne," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, 2008.