

Çok Katmanlı Algılayıcı ve Geriye Yayılım Algoritması ile

Konuşmacı Ayırt Etme

İsmail Aybars Morah, Fırat Fehmi Aygün

Dokuz Eylül Üniversitesi, Bilgisayar Mühendisliği Bölümü, İzmir
ismail.morali@st.cs.deu.edu.tr, firat.aygun@st.cs.deu.edu.tr

Özet: İnsan kulağının konuşmaların sahibini ve şarkı söyleyenleri ayırt etmesi gibi bilgisayar da bu ayrımları yapabilirler. Biz bu çalışmada konuşmacı ayırt etmenin detaylarını inceledik. Geriye yayılım algoritması ve yapay sinir ağları ile verilen ses dosyasının konuşmacısının kim olduğunu tespit etmeye çalıştık. Yapay sinir ağlarının eğitim aşamasından önce MFCC dönüşümü ile önemli ayırt edici özellikleri bulduk. Eğitimden sonra, eğitim verilerini test verileri olarak verdik ve tanıma oranının son ayarlarda %57.5 bulduk.

Anahtar Kelimeler: Konuşmacı tanıma, yapay sinir ağları, geriye yayılım, MFCC.

Identifying Speakers Using Multi Layer Perceptron Backpropagation Algorithm

Abstract: As the human ear and brain identifies speakers of the songs and speeches, the computers can also identify the speakers. In this project, we have looked into the details of speaker identification. It has been aimed to identify the speaker of a given sound file based on the well known backpropagation neural network approach that was used to handle the training and identifying process. Before training phase, the feature set of the training set of sound files were extracted by MFCC transformation. After training, we have given the training set as test set and found the successful identification is 57.5% with final configuration.

Keywords: Speaker identification, neural network, backpropagation, MFCC.

1. Giriş

İnsan sesi çeşitli bilgiler içerir. İsim vermek gerekirse söylenen kelime, söyleme tonu, söyleyenin ruh hali, söyleyenin cinsiyeti, kelimenin dili, söyleyenin kimliği insan sesinden çıkartılabilir. Bu çalışma da Konuşmacı Tanıma'nın bir parçası olan konuşmacıyı ayırt etmeyi sunar.

Konuşmacı tanıma temel olarak iki başlığa ayrılır. Biri konuşmacı doğrulama, diğeri de konuşmacı ayırt etmedir. Konuşmacı doğrulama konuşanın sesinden kişinin kimliğini doğrulama işidir. Bu süreç sadece iddia edilen kimlik hakkında ikili bir karara varmayı içerir.

Aynı şekilde konuşmacı ayırt etme de iki ana başlığa ayrılabilir. Bunlar açık küme ve kapalı küme konuşmacı ayırt etmedir.

Açık küme ayırt etmede amaç, sahibi belli olmayan ses örneğinin hangi kayıtlı konuşmacıya ait olduğunu bulmak veya sesin kimseye ait olmadığını söylemektir. Kapalı küme ayırt etmede ise amaç en çok eşleşen konuşmacıyı bulmaktır [2]. Bu iki yöntem arasındaki tek fark, kapalı küme yöntemi mutlaka bir sonuç döndürür fakat açık küme yöntemi bilinmeyen konuşmacı sonucunu verebilir. Bu çalışmada kapalı küme konuşmacı ayırt etme ele alınmıştır. Kapalı küme konuşmacı ayırt etme de iki dala ayrılır – metin bağımlı ve metin bağımsız.

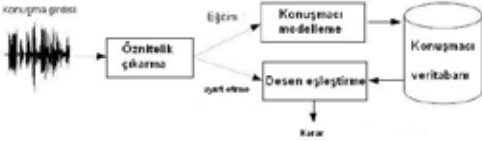
Metin bağımlı yöntemlerde sistem söylenen kelimeyi bilir ve konuşmacıyı bulur. Metin bağımsız yöntemde ise sistem konuşmacıyı herhangi bir metinden bulmak zorundadır ki bu da konuşmacının ayırt edilebilmek için belirli birkaç kelimeyi ya da tümceyi söylemesi zorunluluğunun ortadan kalkması demektir.



Şekil 1: Konuşmacı Tanımanın Sınıflandırılması [2]

2. Kapalı Küme Konuşmacı Ayırt Etme

Tüm desen tanıma işlemleri gibi konuşmacı ayırt etmede de dört ana adım bulunmaktadır (Şekil 2).



Şekil 2 : Konuşmacı Ayırt Etme Sistemi

Konuşma Örnekleri Toplama: Konuşmacı veritabanını oluşturmak için kaydedilmiş olan konuşma örneklerinin toplanmasıdır. Bu bölüm açıkçası basit gözükmemektedir çünkü tek gereksinim kaliteli ve temiz kayıttır ortamıdır.

Fakat, eğer problem konuşmalardan bağımsız olacak ise verilmesi gereken bazı kararlar bulunmaktadır. Konuşulan dilin karakteristik özelliklerinden konuşmacının etkilenmemesi için seslendirilecek kelimeler ve cümlecikler özenle seçilmelidir. Türkçe için biz bir örnek kümesi kararlaştırdık.

Öznitelik Çıkarma: Kaydedilen tüm ses örneklerinin boyutlarının birbirinden farklı olması nedeniyle tüm örneklerin aynı boyutta temsil edilmesi gerekmektedir. Aynı zamanda bazen, ses dosyalarının boyutu, işlenebilmesi için olması gerekenden daha büyük boyutta veya içerikleri pek çok önemsiz veri taşıyor olabilir. Bu sebep ile ses dosyalarını standart hale getirebilmek için her bölümde bir öznitelik çıkartma işlemine ihtiyacımız vardır. Öznitelik çıkartma işlemi ses dosyasından işe yarayan ve istenilen özelliklerinin tespit edilmesini ve onların aynı özellik sayısı ile temsil edilmesini sağlar. Öznitelik çıkartma işlemi konuşmacıların konuşma stilleri arasındaki analitik farkı belirtmelidir. Dış seslerin konuşmalardan ayrılması da önemlidir. Bu tespit edilen özellikler daha sonra konuşmacıların konuşma özellikleri ile karşılaştırılır. Çok farklı öznitelik tespit algoritmaları bulunmaktadır. Mel-Frequency Cepstrum Katsayıları (MFCC), lineer tahminlenebilir kodlama (LPC) (autoregressive modelleme veya AR modelleme olarak da bilinir) konuşmacı tanıma problemlerinde konuşma sinyal özelliklerinin ayırt edilmesinde kullanılan iki yaygın algoritmadır [2], [4].

Konuşmacı Modelleme (Eğitim): Bu adım konuşmacıları konuşma sinyal özniteliklerine göre modellemeyle uğraşır. Amaç konuşmacının sesini eğitimde kullanılacak örneklere göre genelleştirmektir. Diğer bir deyişle görünmeyen yöneyler düzgünce sınıflandırılabilir.

Konuşmacı tanıma sistemlerinde Bayes öğrenmesi ve karar ağaçları kullanılabilir. Eğer Bayes öğrenmesine uygun veri kümesi sağlanabilirse yüksek sınıflandırma oranlarına erişilebilir. Bayes'te öğrenme zamanı yoktur, yani sınıflandırmadan önce de bir başlangıç zamanı gerekmez. Fakat Bayes öğrenmesi her sınıflandırma için tüm veri kümesini işler. Dolayısıyla bu yöntemin ana dezavantajı test verisinin uzun işlem zamanıdır. Bayes yöntemi şifre kontrolü ve sahip tanıma uygulamalarında kullanılabilirama büyük veritabanlarında arama ve sınıflandırmada büyük miktarda zaman harcarlar [3]. Amacımız Napster gibi programlar için

bir sınıflandırıcı yazmak ve iPod gibi avuçiçi cihazlara gömülü programlar yazmaktır. Dolayısıyla Bayes öğrenmesi uygun değildir çünkü Bayes'in kullandığı istatistik yöntem için büyük bir veri kümesine ihtiyaç vardır.

Karar ağaçlarının fonksiyonel yaklaşımı vardır ve sonuç sınıflandırma ağaçları bulurlar. Bu sınıflandırma ağaçlarını bulmak ve ağaçlara göre sınıflandırmak çok zaman almaz [3]. Karar ağacını tercih etmememizin sebebi bu yöntemin yerel en iyi noktasına takılmasıdır. Gömülü bir sisteme zayıf bir sınıflandırıcı vermek tatmin etmeyecektir.

Diğer yöntemlerin arasında konuşmacı tanıma problemleri için arasında sinir ağları temelli modeller popülermiştir çünkü sinir ağları her konuşmacı için yeni bir eğitim yerine, kayıtlı konuşmacıların konuşmalarının farklarını bulacak şekilde eğitilmişlerdir ve bu sayede daha az parametreye ihtiyacı vardır; eğitim ve tanıma bölümlerinde daha iyi performans verir. Yapay sinir ağları bir defa eğitilir ve çalışma zamanları kısadır. Büyük miktarlarda ses dosyalarını (mp3 veritabanları) sanatçıya veya türüne göre sınıflandırmak örnek bir çalışma alanıdır. Bu uygulama hızlı sınıflandırmaya ihtiyaç duyar -işlenecek çok fazla dosya var- ve sinir ağları bu tür işler için uygundur [1].

Desen Eşleştirme: Bu adım aslında bir test adıdır. Tüm konuşmacıların konuşmacı veritabanında nasıl temsil edileceği belirlendikten sonra, bu kayıtlı konuşmacılardan birisinin, yeni bir konuşma örneği alınır. Bu örneğin öznelikleri tespit edildikten sonra bu veriler, veritabanındaki konuşmacıların özellik verileri ile karşılaştırılır. Bilinmeyen özellik yöneyi ve referans şablonunun karşılaştırılmasının ardından, tasarlayan kişinin isteğine göre yanıt olarak, eğer var ise kesin sonuç veya en yakın sonuç döndürülür.

3. Sistem Yapılandırması

3.1. Konuşma Örnekleri Toplama

Kayıt cihazı olarak standart bir mikrofon kullanıldı. Sesler 8.000 Khz, 16 Bit Mono olarak

.WAV dosyası olarak kaydedildi. Tüm kayıt aynı ortamda yapıldı.

3.1.1. Eğitim Kümesi Oluşturma

a. Sözcük Seçimi

Veri kümesindeki kelimeler ve tümceler Türkçe'nin ana sesli harflerine göre ve aralarındaki ilişkiye göre 72 adet seçildi. Bu yolla eğitim kümesinin dilin özelliklerini kapsayacak şekilde dağılması sağlanabilecekti (Tablo 1).

ad	tavşan	fare	elmas	ırlanta
ed	elek	ay	hey kıl	gıybet
id	pırl	adil	deli	kısmi
id	cici	akort	efor	dış borç
od	kokoş	az öl	sektör	kızı öp
öd	öhöm	maymun	eru	kıt us
ud	kusur	akü	eyüp	hırgür
üd	pütür			
dinar	ofsayt	özvatan	kulak	cüzdan
tinsel	obez	özel	kuzen	küpe
iç dış	bozkır	özısı	rus kıızı	yüzyıl
idol	lordi	önbilgi	muzip	müzik
vizör	ondört	göksoy	uzo	büro
ıglu	orsuk	rötuş	rusu öp	güngör
içyüz	bordür	öpücük	uzun ün	vücut

Tablo 1: Seçilen Sözcükler

b. Konuşmacılar

Sistemde 4 konuşmacı kayıtlıdır.

3.1.2. Test Kümesi

Test kümesi olarak eğitim kümesine ek olarak 'laylaylom' ve 'deneme' kelimeleri kullanılmıştır.

3.2. MFCC İle Öznelik Çıkarma

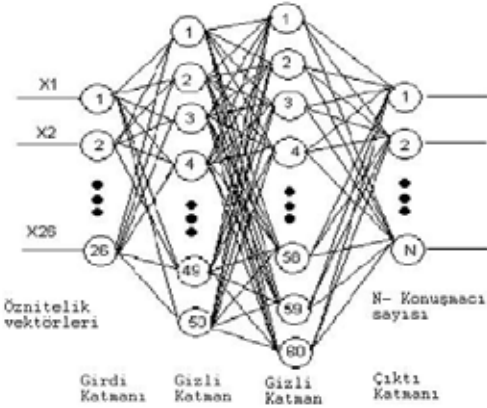
Mel-Frequency Cepstrum Katsayıları, algı temelli sesi temsil eden katsayılarıdır. Fourier Dönüşümü veya Discrete Cosine Dönüşümü'nden türetilir. FFT/DCT ve MFCC arasındaki temel fark MFCC'de frekans bantları logaritmik olarak (Mel ölçüsünde) yerleşmiştir ve bu da insan ses sisteminin yanıtını bantları doğrusal olarak yerleşen FFT veya DCT'ye göre daha da yaklaştırır.

Bir ses dalgasının MFCC'sini bulmak için aşağıdaki yol izlenir:

- Veriyi bir Hamming penceresiyle izle.
- Verinin DFT'sinin genlik değerini bul.
- Genlik değerlerini filtre bankası çıktılarınına çevir.
- 10 tabanında logaritma hesapla.
- Cosine transform'u bul.
- MFCC hesaplarında Malcolm Slaney'in geliştirdiği Auditory Toolbox kullanılmıştır.

3.3. MLP Yapay Sinir Ağı

Konuşmacı ayırt etme problemini çözmek için yapay sinir ağı yapısını kullanmaya karar verdik. Bu seçimi yapmamızın nedenleri arasında, eğitim aşamasının uzun sürmesine rağmen test aşamasının oldukça hızlı gerçekleşmesi ve hata yüzeyi üzerinde minimum hatayı bulmaya çalışırken bölgesel minimumlara takılma oranının çok düşük olması sayılabilir ki her iki sebep de son derece önemlidir. Ayrıca seçtiğimiz çok katmanlı algılayıcılar üzerinde çalışan ileri beslemeli geri yayımlı algoritma, yönetimli bir metot olup, sisteme girdi olarak hangi sesin hangi konuşmacıya ait olduğunu bilgisini de vermemize olanak sağlamaktadır. Sonuç olarak seçilen yapı ve algoritma “bunu söyleyen kim?” sorusunu cevaplandırmaya olanak sağlayan bir yöntemdir.



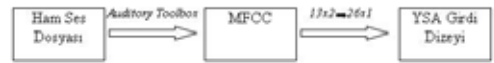
Şekil 3: Konuşmacı tanıma problemi için tipik bir çok katmanlı algılayıcı yapay sinir ağı tasarımı.

Yapay sinir ağını çok katmanlı algılayıcılar üzerinde ileri beslemeli geri yayımlı algoritmayı çalıştıracak şekilde kurguladık. Standart geri

yayımlı algoritması aslında hata yüzeyi üzerinde minimum noktaya ulaşabilmek için eğimli bir inişin öngörüldüğü bir yöntemdir. Bunu sağlayabilmek için de her adımda ağırlık değerleri güncellenir. Öngördüğümüz sınıflandırıcı yapay sinir ağı bir giriş katmanı, iki gizli katman ve bir çıkış katmanı olmak üzere 4 katmandan oluşmaktadır. Giriş katmanındaki nöron sayısı her bir ses dosyasından çıkarılan öznelik sayısına eşit, çıkış katmanındaki konuşmacı sayısına eşit olup, gizli katmanlardaki nöron sayıları bağımsız olarak değiştirilebilmektedir. Tüm katmanlarda etkinleştirme fonksiyonu olarak ‘HyperTansig’ kullanılmıştır. Ayrıca nöronların ilk ağırlıkları da rastsal olarak atanmıştır.

3.4. Eğitim

WAV olarak kaydedilen ses dosyaları okunup “Auditory Toolbox” aracı kullanılarak MFCC ile tanımlanan öznelikler çıkarıldı. Her ses dosyası için 13x2 boyutundaki öznelik dizeyi yapay sinir ağlarına girdi olarak verilebilmesi adına önce doğrusal hale getirildi, 26x1. Sonuç olarak yapay sinir ağlarına girdi olarak verilen dizeyi 26xSöylenen_kelime_sayısı. Söylenen kelime sayısı konuşmacı sayısından bağımsız bir rakam olmakla birlikte iyi sonuç alabilmek adına konuşmacı başına düşen kelime sayıları aynı olmalıdır.



Şekil 4: Veri kümesi ön işleme

Eğitim aşamasında, her adımda, ağırlıklar farklı konuşmacılardan alınan tüm örnekler için ayrı ayrı güncellenir (çevrimiçi güncelleme). Her konuşmacı için bir eğitim kümesi, ve çıkış katmanında özel bir nöron vardır. İşlemin sonunda etkin olan nöron bize konuşmacının kimliğini verecektir. Bir konuşmacı için çıkış katmanındaki ilgili nöron etkin olduğunda diğerleri de pasif kalacaktır. Beklenen çıktılar Konuşmacı_sayısı X 1 boyutundaki bir dizeyi elde edilir. Bu dizeyi bize her konuşmacı için ilgili nöronun aktif değerini verecektir. Örneğin, eğer ses birinci konuşmacı için eğitilmişse, birinci nöronun değeri yüksek diğerlerinin-

ki düşük olacaktır. Bu işlem tüm konuşmacılar ve için baştan belirlenmiş bir adım sayısı kadar tekrar edilir. Son olarak eğitim aşamasından elde edilen ağırlık katsayıları test aşamasında kullanılmak üzere kaydedilir.

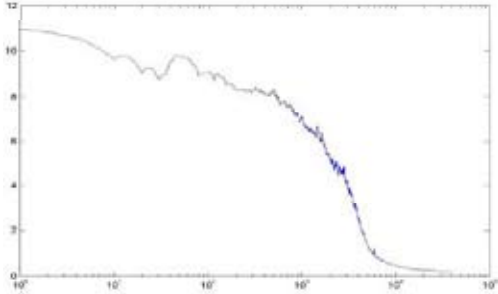
4. Deneyler ve Sonuçlar

Başlangıçta MLP tasarımımızın 4 katmanlı olmasına rağmen, yetersiz hesaplama gücü olanağı nedeniyle, gizli katmanları devre dışı bırakarak iki katmanlı yapı ile çalıştık.

2 katmanlı yapay sinir ağı yapıları üzerinde ileri beslemeli geriye yayılım algoritması kullanıldı. Giriş katmanına 26, çıkış katmanında ise konuşmacı sayısı kadar, 4, nöron yerleştirildi. Üç farklı adım sayısı ile deney yaptık.

4.1. Deney 1

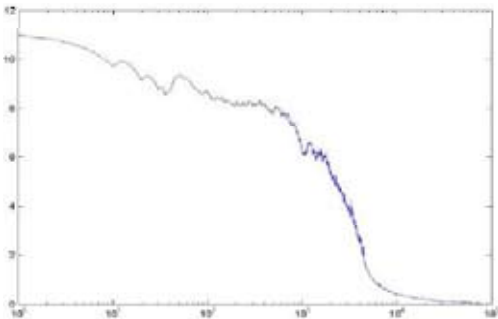
Adım : 1000 ; Ayırt etme oranı: %32,5



Şekil 5: Birinci hata grafiği

4.2. Deney 2

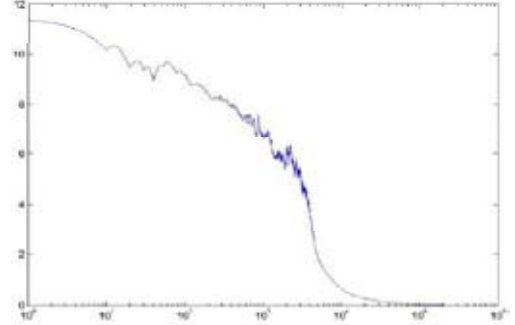
Adım : 2000 ; Ayırt etme oranı: %45



Şekil 6: İkinci hata grafiği

4.3. Deney 3

Adım : 5000 ; Ayırt etme oranı: %57,5



Şekil 7: Üçüncü hata grafiği

5. Sonuçlar

Eğitim aşamasında kullanılan kümeyi test aşamasında da kullandığımızda en iyi ayırt etme oranı %57,5 olarak gerçekleşti. Aslına bakılacak olursa, sonuç beklenildiği kadar doğru olmamakla birlikte, bunun belirgin bazı sebepleri de vardır.

Verilen 3 farklı deney sonucundan da görülebileceği gibi, adım sayısı ile ayırt etme oranı arasında doğrusal bir ilişki var. Bu ilişkiye dayanarak, 4 katmanlı MLP yapısı kullanıldığında ve adım sayısı artırıldığında, şüphesiz ki çok daha iyi sonuçlar alınabilecektir.

Yine yapısal bir iyileştirme olarak nöron ağırlıklarının ilk değerlerinin atanmasında, nöronları hata yüzeyine düzgün dağıtacak farklı bir ilkleme işlevi kullanılabilir.

Ayrıca eğitim kümesinin dilin özelliklerine dikkat edilerek genişletilmesi de, bu küme dışından seçilen kelimeler için yapılacak testlerde başarı oranını arttıracaktır.

Kayıt ortamının iyileştirilmesi, yani kullanılan mikrofonun kalitesinin artırılması, ve ortam dış seslerden yalıtılması da kaydedilen örneklerden çıkarılan özniteliklerin daha ayırıcı olmasını sağlayacaktır.

6. Kısaltmalar

FFT	Hızlı Fourier Dönüşümü
DFT	Ayrık Fourier Dönüşümü
MLP	Çok katmanlı algılayıcı

7. Teşekkürler

Yöney nicemleme işlemlerindeki ve Malcom Slanley'in MFCC fonksiyonunu düzenleme-
deki yardımlarından ötürü Dr. Emine Ekin'e
teşekkür ederiz.

8. Kaynaklar

[1]. Ham F.M., Kostanic I. *Principles of Neurocomputing for science and Engineering*, McGraw-Hill, ISBN: 0070259666, 2001.

[2]. Karpov E., Real-Time Speaker Identification, Master's Thesis, University of Joensuu, Department of Computer Science, 2003.

[3]. Mitchell T.M., *Machine Learning*, McGraw-Hill, ISBN: 0070428077, 1997.

[4]. Sharma A., Singh P.S., Kumar V., "*Text-independent speaker identification using back-propagation MLP network classifier for a closed set of speakers*", Proceedings of 2005 IEEE International Symposium on Signal Processing and Information Technology, 665-669, 2005.