Konuşmacı Tanıma Sistemi için Yeni Bir Sınıflandırıcı
A New Classifier for Speaker Identification System

Figen ERTAŞ[1], Cemal HANİLÇİ[2]

[1]Elektronik Mühendisliği
Uludağ Üniversitesi
fertas@uludag.edu.tr, chanilci@uludag.edu.tr

Özet

Bu bildiride, metinden bağımsız konuşmacı tanıma sistemi için yeni bir sınıflandırıcı, sınıfsal temel olmayan bileşen analizi (STOBA), TIMIT veritabanı için incelenmiş olup tanıma başarımı ve zaman açısından bilinen en yaygın sınıflandırıcılardan biri olan Vektör Nicemleme algoritması ile karşılaştırılmıştır. VN ile 220 saniyede %100 başarım elde edilirken, STOBA sınıflandırıcısı ile 40 saniyede %98.67 tanıma başarımı elde edilmiştir.

Abstract

In this paper, a new classifier for text-independent speaker identification system, Classwise Non-principal Component Analysis (CNPCA), has been analyzed and compared with Vector Quantization (VQ) classifier which is one of the most common classifier, in relation to recognition rate and training and testing duration using TIMIT database. However 100% identification rate has been obtained in 220 seconds by using VQ classifier, 98.67% identification rate has been obtained by CNPCA in 40 seconds.

1. Giriş

Konuşmacı tanıma, ses işaretinin içerdiği bilginin kullanılması ile otomatik olarak kimin konuştuğunun belirlenmesi işlemidir. Konuşmacı tanıma, konuşmacı doğrulama ve konuşmacı belirleme olmak üzere ikiye ayrılır. Konuşmacı doğrulama, verilen bir ses örneğinin iddia edilen kişiye ait olup olmadığının tespiti, konuşmacı belirleme ise verilen ses örneğinin sistemde daha önceden kayıtlı olan kişilerden hangisine ait olduğunun saptanması işlemidir [1], [2]. Ayrıca konuşmacı tanıma işlemleri metinden bağımsız veya metne bağımlı olarak da iki gruba ayrılabilir. Metinden bağımsız konuşmacı tanıma sistemlerinde sistemin eğitim ve test aşamalarında farklı cümleler veya sözcükler kullanılırken metne bağımlı sistemlerde hem eğitim hem de test sırasında aynı cümle veya sözcükler kullanılmaktadır. Konuşmacı tanıma sistemlerinin kullanım alanları oldukça yaygındır. Örneğin son yıllarda, telefon bankacılığı, sesli arama, telefonla alışveriş, veritabanı erişim servisleri, bilgisayarların uzaktan sesle kontrolü ve en önemlilerinden biri de adli uygulamalar gibi birçok alanda kullanılmaya başlanmıştır [3].

Ses işaretinin zamanla çok sık değişmesi (durağan olmaması), ortamdaki gürültü ve hava şartları gibi etmenlerden kolayca etkilenmesi nedeniyle bir örüntü tanıma problemi olarak konuşmacı tanıma işlemi diğer problemlere nazaran daha zor bir uygulamadır. Ses işaretinin durağan olmaması sistem tasarımcılarını işareti kısa zaman aralıklarında (10-20 ms) işlemeye zorlamakta ve bu zorunluluk veri boyutlarının oldukça yüksek olmasına neden olmaktadır [4], [5]. Büyük veri boyutları ise sınıflandırıcıların çalışma sürelerini artırmaktadır. Bundan dolayı bir konuşmacı tanıma sisteminde kullanılan sınıflandırma yönteminin hem yüksek başarım vermesi hem de hızlı sonuç üretmesi beklenmektedir.

Bu çalışmada metinden bağımsız konuşmacı tanıma problemlerinde yaygın olarak kullanılan ve mevcut sınıflandırıcılar arasında en hızlı sonuç üreten sınıflandırma algoritması olan Vektör Nicemleme (VN) [6], [7] algoritması ile yeni bir sınıflandırıcı algoritma, sınıfsal temel olmayan bileşen analizi (Classwise Non-Principal Component Analysis), STOBA [8], tanıma başarımı ve zaman açısından karşılaştırılmıştır.

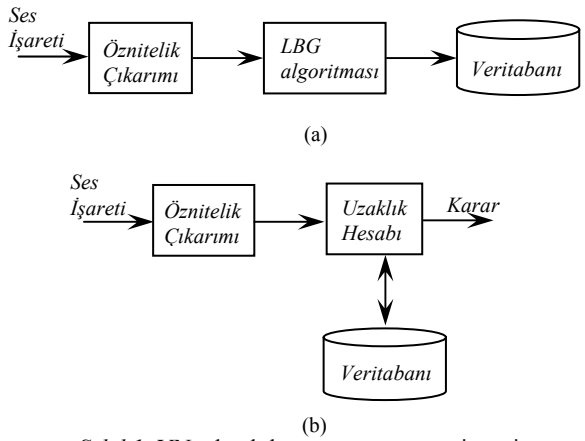
2. Yöntem

2.1. Vektör Nicemleme

Şekil 1'de Vektör Nicemleme tabanlı konuşmacı tanıma sisteminin genel yapısı gösterilmiştir. *Öznitelik çıkarımı,* ses işaretinden kişiyi temsil parametrelerin elde edilmesi işlemidir. En çok kullanılan öznitelikler Mel Frekansı Kepstrum Katsayıları (MFCC) ve Doğrusal Öngörülü Kodlama Katsayılarıdır (LPCC) [4], [5]. Öznitelikler, ses işaretinin kısa-dönem karakteristiğini ve ses yolunun özelliklerini içerir [7].

Eğitim aşamasında, her konuşmacı için eğitim verileri kullanılarak matematiksel bir model (kod kitabı) oluşturulur ve oluşturulan model veritabanında saklanır. LBG algoritması [6] uygulanabilirliğinin basit ve etkili olması nedeniyle kod kitabı oluşturulmasında kullanılan en yaygın yöntemlerden biridir. Eğitim aşamasında ses işaretinden elde edilen öznitelik vektörleri LBG algoritması ile vektör uzayında kümelere ayrılır ve aynı zamanda veri miktarı da bu sayede azaltılmış olur. LBG algoritmasının girişi öznitelik vektörleri X, çıkışı ise oluşturulan kümelerin ağırlık merkezlerini (kod vektörleri) içeren kod kitabıdır (C). Kod kitabı her konuşmacıya ait

modeli temsil etmektedir. Kod kitabı oluşturulmasındaki amaç M adet vektörün N adet vektör ile temsil edilmesidir *(N<<M)* .



Şekil 1: VN tabanlı konuşmacı tanıma sistemi
(a) Eğitim Aşaması (b) Test Aşaması

VN algoritması ile konuşmacı tanıma sisteminin test aşaması ise şu adımlardan oluşur:

Adım 1: Test işareti için öznitelik vektörleri elde edilir, $X = \{x_i\}$

Adım 2: Veritabanında kayıtlı her bir konuşmacı modeli (kod kitabı) C_i için X ve C_i arasındaki bozulma miktarı, $D_i = d(X, C_i)$, hesaplanır

Adım 3: Bilinmeyen ses işareti bozulma miktarı en küçük olan modele atanır.

$$Id = \arg\min_{i=1,2,...,K} \{D_i\} \quad (1)$$

Adım 2'de hesaplanan bozulma miktarı, kod kitabı C_i, ile öznitelik vektörü, X, arasındaki benzerliğin bir ölçüsüdür. Bu çalışmada bozulma miktarının hesaplanmasında sıklıkla kullanılan ve denklem (2) ile hesaplanan metrik kullanılmıştır.

$$d(X, C_i) = \frac{1}{L} \sum_{j=1}^{L} \min_{k=1}^{K} d_E(x_j, c_{ik}) \quad (2)$$

Denklem (2) deki *L* özellik vektörü boyutunu, *K* ise veritabanında kayıtlı bulunan konuşmacı sayısını belirtmektedir. Yani bozulma miktarı hesaplanırken *X* kümesindeki her bir vektör *C* kod kitabı içerisindeki en yakın kod vektörüne atanır ve bütün uzaklıklar üzerinden ortalama alınır. Denklem (2) de belirtilen d_E iki vektör arasındaki Öklit mesafesidir ve denklem (3) ile hesaplanır.

$$d_E(x, y) = \sqrt{\sum_{i=1}^{M} (x_i - y_i)^2} \quad (3)$$

2.2. Sınıfsal Temel Olmayan Bileşen Analizi (STOBA)

Ses işaretinden elde edilen öznitelik vektör boyutlarının yüksek olması sebebi ile sınıflandırıcı algoritmaların sonuç üretme süresi artmaktadır. Yüksek boyutlu verilerle ilgili bu problemin çözümü için yaygın olarak kullanılan yöntemlerden birisi temel bileşen analizidir (TBA). TBA ile yüksek boyutlu veriler daha düşük boyutlu verilere indirgenebilmektedir. Ancak TBA sınıflandırma probleminde genellikle tatmin edici sonuç vermemektedir. TBA nın bu sınırlamasından dolayı Xuan ve arkadaşları tarafından yeni bir sınıflandırıcı, STBOA, geliştirilmiştir [8]. STOBA, her sınıf için sınıf ortalaması ve sınıf değişintisi gibi karakteristikleri kullanmaktadır. STOBA, şu şekilde özetlenebilir:

X k. sınıfa ait ses eğitim verilerinden elde edilen *n* boyutlu öznitelik vektörleri olsun,

$$X = [x_1, x_2, ...]$$

Öznitelik vektörlerinin ortalaması, $\overline{X}$, hesaplandıktan sonra $\hat{X}$ fark matrisi hesaplanır.

$$\hat{X} = [x_1 - \overline{X}, x_2 - \overline{X}, ...]$$

Fark matrisinden denklem (4) ile ortak değişinti matrisi elde edilir.

$$C = \frac{1}{n} \hat{X} \hat{X}^T \quad (4)$$

Ortak değişinti matrisinden Φ_k özvektörleri ve Λ_k özdeğer matrisi elde edilir. Özdeğerler büyükten küçüğe doğru azalan şekilde sıralandığında buna karşılık gelen özvektör matrisi Φ_k denklem (5)'de belirtildiği şekilde ifade edilir:

$$\Phi_k = (\Phi_k)_{n \times n} = [\Phi_{rk}, \Psi_{rk}]_{n \times n} \quad (5)$$

Denklem (5)'de belirtilen $\Phi_k = (\Phi_k)_{n \times n}$, *k.* sınıfa ait eğitim verilerinden elde edilen özvektörlerin tamamının oluşturmuş olduğu özvektör matrisidir. $\Phi_{rk} = (\Phi_{rk})_{n \times r}$, *k.* sınıfa ait *r* adet özvektöre karşılık gelen temel bileşen matrisi iken $\Psi_{rk} = (\Psi_{rk})_{n \times (n-r)}$ ise *(n-r)* adet özvektöre karşılık gelen temel olmayan bileşen matrisidir. Kısaca, denklem (5)'de belirtilen *n* değeri öznitelik vektörü boyutunu, *r* $(r \le n)$, en yüksek değere sahip özdeğere karşılık gelen özvektör sayısı, *(n-r)* ise *r* adet en yüksek özvektör haricinde kalan özvektör sayısını belirtmektedir. Bu bilgiler ışığında STOBA ile sınıflandırma işlemi şu adımlardan oluşur:

Adım 1: Her sınıf için verilen eğitim verilerini kullanarak özdeğer ve özvektörler hesaplanır. *(n-r)* adet en küçük özdeğerlere karşılık gelen, Ψ_{rk}, özvektör matrisi elde edilir.

Adım 2: Her sınıf için M_k, ortalama vektörü hesaplanır.

Adım 3: Sistemin girişine uygulanan test örneği, y, için denklem (6) ile belirtilen her sınıf ve test örneği arasındaki STOBA uzaklığı hesaplanır.

$$D_{rk} = \left\| \Psi_{rk}^T (y - M_k) \right\| = (y - M_k) \Psi_{rk} \Psi_{rk}^T (y - M_k) \quad (6)$$

Adım 4: Bilinmeyen ses örneği y, minimum uzaklık veren sınıfa atanır.

$$\hat{k} = \arg\min_{k}(D_{rk})$$

3. Deneysel Çalışma

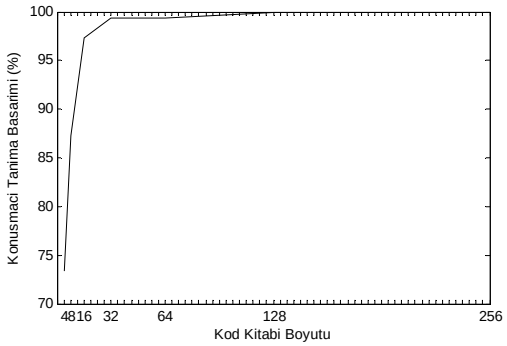
Konuşmacı tanıma deneylerinde TIMIT veritabanı kullanılmıştır. TIMIT veritabanı Amerikan İngilizcesinin 8 ana lehçesini konuşan, 438'i erkek 192'si kadın olmak üzere toplam 630 konuşmacının her birinin fonetik yönden zengin 10'ar adet cümlesini içerir [9].

Bu çalışmada, öznitelik vektörü olarak mel ölçekli kepstrum katsayıları kullanılmıştır [10]. Ses işareti, 10 ms'lik kısmı örtüşen 20 ms uzunluğundaki çerçevelere ayrılıp Hamming pencere uygulanarak işlenmektedir. Pencerelenen ses işaretinin 512 örnek uzunluklu Hızlı Fourier Dönüşümü (HFD) alınıp, elde edilen vektör mel ölçekte 0-8000 Hz arasına yerleştirilmiş üçgen süzgeç takımına uygulanmıştır. Her bir çerçeveye karşılık olarak TIMIT veritabanı için 24 boyutlu öznitelik vektörleri elde edilmiştir.

Konuşmacı tanıma deneylerinde veritabanlarının test dizinine ait 50 konuşmacı (25 erkek, 25 bayan) kullanılmıştır. VN ile konuşmacı tanıma sistemi için değişik kod kitabı boyutları ile tanıma başarımı elde edilmiştir. STOBA sınıflandırıcısı için ise değişik *r* değerlerine ilişkin tanıma başarımları elde edilmiştir. Deneyler INTEL Q6600 işlemci ve 4 GB hafızaya sahip bir bilgisayarda MATLAB 7.0 ortamında gerçekleştirilmiştir.

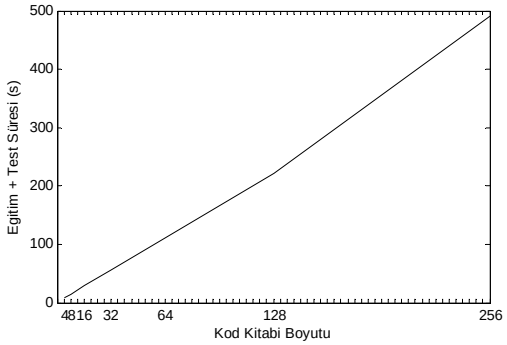
3.1. VN ile Konuşmacı Tanıma

VN ile konuşmacı tanıma deneylerinin ilk aşamasında TIMIT veritabanından alınan 50 konuşmacı için değişik kod kitabı boyutlarına ilişkin tanıma başarımları elde edilmiştir. Her konuşmacı 7 adet cümlesi ile eğitilmiş kalan 3 cümle ile test edilmiştir. Testler ayrık yapılmıştır (50 konuşmacı için toplam 150 test). TIMIT veritabanı için tanıma başarımının kod kitabı boyutuna bağlı değişimi Şekil 1'de gösterilmiştir.



Şekil 1: TIMIT veritabanı için tanıma başarımının kod kitabı boyutu ile değişimi

Kod kitabı boyutuna bağlı olarak sistemin çalışma süresinin değişimi Şekil 2'de gösterilmiştir.

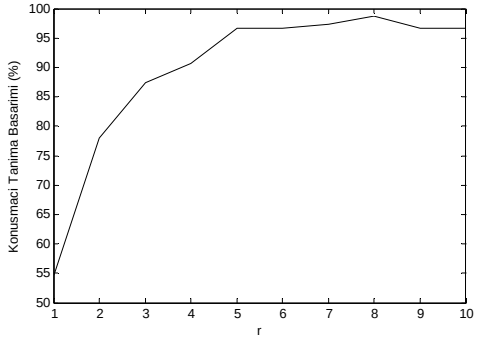


Şekil 2: Kod kitabı boyutunun sistemin çalışma süresine (Eğitim+Test Süresi) etkisi

Şekil 1'den görüldüğü gibi VN ile konuşmacı tanıma sisteminde kod kitabı boyutunun artması tanıma başarımını da artırmaktadır. VN ile yapılan testlerde maksimum başarım kod kitabı boyutunun 128 ve 256 olarak seçildiği durumlarda elde edilmiştir (%100). Bununla birlikte kod kitabı boyutu arttıkça sistemin çalışma süresi de artmaktadır (Şekil 2).

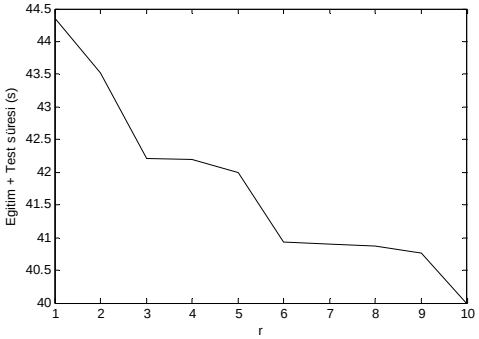
3.2. STOBA ile Konuşmacı Tanıma

Deneysel çalışmanın ikinci aşamasında konuşmacı tanıma sisteminde STOBA sınıflandırıcısı kullanılmıştır. Şekil 3'de STOBA sınıflandırıcısı kullanılarak TIMIT veritabanı için konuşmacı tanıma başarımının *r* parametresine bağlı değişimi verilmiştir. Deneylerde *r* parametresi 1-23 arasında değiştirilmiş ve elde edilen sonuçlarda maksimum başarım *r=8* için elde edilmiştir (%98.67). Şekil 4'de ise *r* parametresindeki değişimin sistemin çalışma süresine etkisi gösterilmiştir. Şekil 3 ve 4'den de görüldüğü üzere STOBA sınıflandırıcısı VN kadar yüksek başarım verirken aynı zamanda da VN sınıflandırıcısına göre oldukça hızlı bir algoritmadır. STOBA sınıflandırma yaparken Bölüm 2.2'de de anlatıldığı gibi aynı zamanda boyut indirgeme işlemi yapmaktadır. Verilerin tamamını kullanmak yerine özvektörlerden oluşan matrisi kullanarak boyutta bir indirgeme yapmaktadır.



Şekil 3: Konuşmacı tanıma başarımının *r* parametresine bağlı değişimi

Şekil 4'den de görüldüğü gibi STOBA çok hızlı bir algoritmadır ve *r* parametresi arttıkça sistemin çalışma süresi azalmaktadır. Çünkü *r* parametresi arttıkça sınıflandırmada kullanılan ve denklem (6)'da belirtilen özvektörlerden oluşan matris boyutu azalmaktadır.



Şekil 4: r parametresinin sistemin çalışma süresine etkisi

4. Sonuç

Bu çalışmada metinden bağımsız konuşmacı tanıma problemi için yeni bir sınıflandırıcı algoritma olan STOBA, bilinen en yaygın ve en hızlı sınıflandırıcı algoritmalardan biri olan VN algoritması ile tanıma başarımı ve sistemin çalışma süresi açısından karşılaştırılmıştır. Elde edilen sonuçlardan da görüleceği gibi STOBA uygulanabilirliği, modellenmesi ve programlanabilirliği açısından oldukça kolay bir algoritma olup VN algoritmasından daha kısa zamanda sonuç üretmektedir. STOBA ile maksimum tanıma başarımı yaklaşık 40 saniyede elde edilirken, VN ile maksimum tanıma başarımı 220 saniyede elde edilmiştir. Tanıma başarımı açısından ise, en uygun parametreler ile VN algoritması %100 tanıma başarımı verirken STOBA ile %98.67 tanıma başarımı elde edilmiştir.

5. Kaynaklar

[1] Doddington, G. R. "Speaker Recognition-Identifying People by Their Voices", *Proceedings of The IEEE*, 73, 11, 1651-1665, 1985

[2] Campbell, J. P. "Speaker Recognition: A Tutorial", *Proceedings of The IEEE*, 85, 9, 1437-1997, 1997

[3] Furui, S. "Recent Advances in Speaker Recognition", *Pattern Recognition Letters*, 18, 859-872, 1997.

[4] Rabiner, L. R. ve Juang, B. H. *Fundamentals of Speech Recognition*, Prentice Hall, New Jersey, 1993

[5] Deller, J. R., Proakis, J. G. ve Hansen, J. L., *Discrete-Time Processing of Speech Signals*, Macmillan Publishing Company, New York, 1993

[6] Linde, Y., Buzo, A. ve Gray, R. "An Algorithm for Vector Quantization", *IEEE Trans. on Communications*, 28, 1, 84-95, 1980

[7] Kinnunen, T., Karpov, E. ve Franti, P. "Real-Time Speaker Identification and Verification", IEEE Trans. on Audio, Speech, And Language Processing, 14, 1, 277-288, 2006

[8] Xuan, G., Chai, P., Zhu, X., Shi, Y. Q. ve Fu, D. "A Novel Pattern Classification Scheme: Classwise Non-Principal Component Analysis (CNPCA)", *IEEE International Conference on Pattern Recognition,* 3, 320-323, 2006

[9] Reynolds, D. A., Zissman, M. A., Quatieri, T. F., O'Leary, G. C. ve Carlson, B. A. "The effects of telephone transmission degradations on speaker recognition performance", *IEEE International Conferance on Acoustics, Speech and Signal Processing, ICASSP*, 1, 9, 329-332, 1995

[10] Reynolds, D. A., "Gaussian Mixture Modeling Approach to Text-Independent Speaker Identification", *Ph. D. Thesis, Georgia Institute of Technology,* 1992