

Tümleşik Gauss Modeller ve Vektör Nicemleme Yöntemleriyle Konuşmacı Tanıma

⁺Özgür Devrim Orman, *Levent M. Arslan

⁺TÜBİTAK

Ulusal Elektronik ve Kriptoloji Araştırma Enstitüsü, Gebze, 41470 Kocaeli

oorman@mam.gov.tr

*Boğaziçi Üniversitesi

Elektrik-Elektronik Mühendisliği Bölümü, Bebek, 80815 İstanbul

arslanle@boun.edu.tr

Özetçe

Bu çalışmada TIMIT [1] veri tabanında bulunan sekiz farklı lehçedeki tüm konuşmacıları kapsayan bir test öbeği üzerinde Tümleşik Gauss Modeller (TGM) ve Vektör Nicemleme (VN) yöntemleri kullanılarak metinden bağımsız konuşmacı tanıma uygulamaları geliştirilmiştir. Sınama sonuçlarından TGM'in VN yöntemine göre daha iyi bir başarımla verildiği görülmektedir. Konuşmacı tanıma uygulamalarında kullanılan ses nitelik kümesinin seçimi başarımla doğrudan etkilemektedir. Konuşmacı tanıma uygulamaları üzerindeki çalışmalar gelecekte de sürekli gelişen bir çizgide varlığını sürdürecektir

1. Giriş

Konuşmacı tanıma uygulamaları kişisel iletişim olanaklarının gelişimine paralel olarak her geçen gün daha da önem kazanmaktadır. Örnek verecek olursak, telefon ve İnternet üzerindeki bankacılık hizmetlerinin sunumunda kullanıcı doğrulanırken kişiye özgü ses niteliklerinin denetimi sistem güvenliğini attıran bir etmendir. Öte yandan insanoğlunun her gün defalarca başvurduğu bu yeteneğin bilgisayar ortamında benzetimini yapmak için kullanılması gereken en iyi yöntem ve en iyi ses nitelik kümesi henüz belirlenmemiştir.

Konuşmacı tanıma uygulamalarında hem VN yönteminin hem de TGM'nin başarımla önceki çalışmalarda incelenmiştir [3, 4]. Söz konusu iki yöntemin dışında Saklı Markov Modeller (SMM) [5] ve Yapay Sinir Ağları (YSA) [6] da konuşmacı tanıma uygulamalarında kullanılmaktadır. Bu çalışmada TGM ve VN yöntemlerinin TIMIT ses veri tabanında kayıtlı bulunan konuşmacılar üzerindeki metinden bağımsız kapalı küme konuşmacı tanıma başarımları karşılaştırılmıştır. İkinci bölümde konuşmacı tanıma metodolojisinden söz edilmektedir. Sonuç bölümünde ise bu çalışma ile ilgili bilgilere ek olarak yazarların aynı konu üzerindeki sürmekte olan çalışmalarına da değinilmiştir.

2. Metodoloji

Metinden bağımsız kapalı küme konuşmacı tanıma uygulamaları üç aşamaya indirgenebilir, bunlar; seçilen bir dönüşümle kişiye ait konuşma kayıtlarından kişisel ses niteliklerini ifade eden vektörel bir veri öbeği oluşturma, belirlenmiş olan bir yöntem ile bu veri öbeğinin işlenmesi sonucunda kişinin sisteme tanıtılması ve son olarak da bağımsız bir metinden oluşturulmuş kişisel veri öbeği ile sistemin kişiyi tanıma yetisinin sınanması.

Kişisel ses niteliklerinin ifadesinde farklı dönüşümlerden oluşturulabilen kepsral katsayılar kullanılabilir. Bu çalışmada sıklık güç dağılımından elde edilen iki farklı kepsral katsayı türü üzerinde durulmuştur. Yukarıda değinilen aşamalardan ilki olan kişisel ses niteliklerini ifade eden vektörel bir veri öbeği oluşturma süreci maddeler halinde şu şekilde ifade edilebilir:

- Kişiye ait ses kayıtlarının zaman düzleminde birbiriyle örtüşen dizgeler halinde bölütlenmesi.
- Her bir bölütün pencere işlevi ile çarpılması (Hamming penceresi kullanılmıştır).
- Pencerelenmiş bölütlerin hızlı fourier dönüşümü kullanılarak sıklık düzlemine aktarımı.
- Sıklık düzlemi güç dağılımlarının hesaplanması.
- Kepstral katsayıların belirlenmesi (log-kepstrum ve mel-kepstrum).

Oluşturulan veri öbeği tanıtım ve sınama olmak üzere iki farklı kısımdan oluşur. Veri öbeğinin iki farklı kısmındaki yöneyleri $\underline{x}_{i,j,T}$ ve $\underline{x}_{i,j,S}$ olarak ifade edebiliriz, burada i konuşmacıyı, j yöneyin sıra numarasını, T,S ise yöneyin tanıtım veya sınama gurubunda olduğunu göstermektedir.

Konuşmacı tanıma uygulamasının genel ifadesindeki diğer iki aşama olan tanıtım ve sınama süreçleri uygulamada seçilmiş yönteme bağlı olarak farklı şekillerde yürütülürler. Her bir konuşmacı için M boyutlu N adet kod yöneyi içeren VN yöntemi ile oluşturulmuş bir konuşmacı tanıma sisteminde söz konusu süreçleri aşağıdaki aşamalara bölebiliriz.

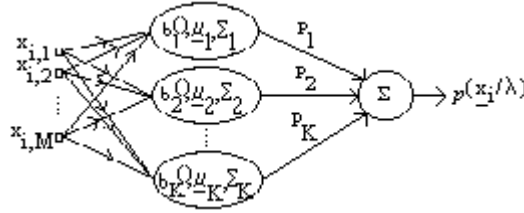
- Kod Kitabının Oluşturulması**

Her konuşmacı için konuşmacıya ait tanıtım yöneyleri Linde-Buzo-Gray yöntemiyle [2] N adet kod yöneyine dönüştürülür.

- Sınama:**

Sınanacak konuşmacının sınama öbeğindeki her bir yöney tüm konuşmacıların kod kitaplarındaki yöneylerle uzaklık-yakınlık ölçütüne göre karşılaştırılır. Sınanan yöneyin en yakın olduğu yöneye sahip olan konuşmacıya ait olduğu varsayılır. Bu şekilde konuşmacının sınama öbeğindeki her bir yöneyin ait olduğu konuşmacı ya da konuşmacılardan en fazla yöneye sahip olan, sınama öbeğinin sahibi olan konuşmacı olarak kabul edilir.

Diğer bir yöntem olan TGM ise yukarıdakinden farklı bir süreç dizisi içerir. Bu yöntemin temel prensibi kişiye ait tanıtım yöneylerinden kişisel akustik niteliklerin olasılıksal yoğunluk işlevinin birden çok Gauss yoğunluk işlevi kullanılarak gösterimidir. Bu yöntemde her bir konuşmacı Sekil 1’de görüleceği gibi K adet Gauss yoğunluk işlevi ile tanımlanır.



Şekil 1: Bir konuşmacıya ait tüm Gauss yoğunluk işlevleri.

Konuşmacı yoğunluk işlevlerinin ortalama, saçılım ve ağırlık değerleri i . konuşmacı için aşağıdaki gibi gösterilir. Bir konuşmacıya ait K adet ağırlık değerlerinin toplamı “1” olmak zorundadır.

$$\lambda_i = \{p_j, \underline{\mu}_j, \Sigma_j\} \quad j = 1, \dots, K$$

$\underline{\mu}_j$: j . Gauss işlevinin ortalama yöneyini,

Σ_j : j . Gauss işlevinin saçılım matrisini,

p_j : j . Gauss işlevinin toplamdaki ağırlık değerini göstermektedir.

Sınama kümesindeki herhangi bir yöneyin bir konuşmacıya ait olma olasılığı söz konusu yöneyin her bir Gauss işlevindeki yoğunluk değerlerinin ağırlıklı toplamıdır. Yöneylerin M boyutlu olması varsayımı altında bu olasılık değerini şu şekilde ifade edebiliriz:

$$p(\underline{x}/\lambda_i) = \sum_{j=1}^K p_j b_j(\underline{x})$$

$$b_j(\underline{x}) = \frac{1}{(2\pi)^{M/2} |\Sigma_j|^{1/2}} \exp\left\{-\frac{1}{2} (\underline{x} - \underline{\mu}_j)^T \Sigma_j^{-1} (\underline{x} - \underline{\mu}_j)\right\}$$

Konuşmacıya ait Gauss yoğunluk işlevlerinin konuşmacıyı ifade edebilmesi için konuşmacıya ait tanıtım yöneyleri kullanılarak Beklenen Ençoklanması (BE) yöntemiyle her bir işlev ayrı ayrı güncellenir. Tanıtım öbeğinde D adet yöney bulunan i . konuşmacı için BE yönteminin uygulanması aşağıdaki gibi ifade edilebilir.

a. Her bir yoğunluk işlevi için yöneylerin soncul olasılıklarının hesaplanması.:

$$p(r/\underline{x}_{i,j,T}) = \frac{p_r b_r(\underline{x}_{i,j,T})}{\sum_{k=1}^K p_k b_k(\underline{x}_{i,j,T})} \quad r = 1, \dots, K$$

b. Ortalama yöneylerinin güncellenmesi:

$$\underline{\mu}_k = \frac{\sum_{j=1}^D p(k/\underline{x}_{i,j,T}, \lambda_i) \underline{x}_{i,j,T}}{\sum_{j=1}^D p(k/\underline{x}_{i,j,T}, \lambda_i)}$$

c. Saçınım matrisinin elemanlarının güncellenmesi:

$$\sigma_k^2 = \frac{\sum_{j=1}^D p(k/\underline{x}_{i,j,T}, \lambda_i) \underline{x}_{i,j,T}^2}{\sum_{j=1}^D p(k/\underline{x}_{i,j,T}, \lambda_i)}$$

d. Ağırlıkların güncellenmesi:

$$p_k = \frac{1}{D} \sum_{j=1}^D p(k/\underline{x}_{i,j,T}, \lambda_i)$$

Güncelleme işlemi konuşmacıya ait tüm Gauss yoğunluk işlevleri konuşmacıyı yeterli doğrulukta ifade edene kadar sürdürülür.

Herhangi bir konuşmacının TGM yönteminde sınanması iki aşamadan oluşmaktadır. İlk aşamada söz konusu konuşmacıya ait sınama yöneylerinin tamamının tanıtılmış tüm konuşmacıların her birine ait olma olasılıkları hesaplanır. İkinci aşamada ise bulunan bu olasılıkların en büyüğüne sahip konuşmacı sınama kümesine sahip konuşmacı olarak belirlenir. Belirlenen konuşmacıyı H ile sınanan konuşmacıya ait tüm sınama yöneylerini de X_S ile gösterirsek, konuşmacının belirlenmesi sürecini şu şekilde ifade edebiliriz:

$$H = \arg \max_{1 \leq i \leq I} \Pr(\lambda_i / X_S) \quad (1)$$

Bayes kuralını gözönüne aldığımızda $\Pr(\lambda_i / X_S)$ şu şekilde yazılabilir:

$$\Pr(\lambda_i / X_S) = \frac{p(X_S / \lambda_i) \Pr(\lambda_i)}{p(X_S)}$$

Tüm konuşmacıların eşit olasılıklı olması ve $p(X_S)$ değerinin de tüm konuşmacılar için aynı olması varsayıldığında bu değerler karşılaştırma açısından bir özellik içermezler. Denklem 1 yeniden yazılacak olursa, şu ifade elde edilir:

$$H = \arg \max_{1 \leq i \leq I} p(X_S / \lambda_i)$$

3. Uygulamalar

İlk aşama olan ses nitelik veri öbeklerinin oluşumunda; tanıtım öbekleri, TIMIT veri tabanındaki her konuşmacının dizininde bulunan “sa” ve “si” önekli ses kayıt dosyalarından, sınama öbekleri de “sx” önekli ses kayıt dosyalarından elde edilmiştir. TIMIT ses veri tabanında “sa” önekli ses kayıt dosyalarının içerdikleri cümleler tüm konuşmacılar için aynıdır.

Vektör Nicemeleme yönteminde aynı uzunlukta vektörler kullanılarak yapılan sınamalarda mel-sıklık kepsstral katsayılarının diğer yöntemlere göre daha yüksek bir başarımla elde edildiği gözlenmiştir. Başarımlardaki farklılığın nedeni, her iki yöntemden elde edilen aynı uzunluklu vektörler karşılaştırıldığında mel-sıklık kepsstral katsayılarının sıklık düzleminde daha geniş bir bölge kaplamasıdır. Tümleşik Gauss Modellerin uygulanmasında yalnızca mel-sıklık kepsstral katsayıları kullanılmıştır.

Bu çalışmada VM yöntemi ile konuşmacı tanıma gerçekleştirilirken kişisel kod kitaplarının her biri 32 katsayıdan oluşan 32 adet vektör içerecek şekilde oluşturulmuştur. TGM yöntemi çalışmada uygulandığında

tanıtım ve sına ma öbekleri 32 boyutlu vektörlerden oluşturulmuştur, her bir konuşmacı da 32 adet Gauss yoğunluk işleviyle ifade edilmiştir. Genel olarak seçilen Gauss yoğunluk işlevinin adedinin artımı başarımdaki iyileşmeyi de beraberinde getirmektedir. Öte yandan eklenen her işlev hem güncelleme sürecinde hem de sına ma sürecinde ek bir işlemsel yük getireceği için Gauss yoğunluk işlevi sayısı uygulamada bu gibi etmenler de göz önüne alınarak belirlenmelidir.

Tüm sına ma öbeklerinin (462 konuşmacı) denenmesi sonucunda VN yönteminde %99,35'lik ve TGM yönteminde de %99,56'lık bir başarı m gözlenmiştir.

4. Sonuç

İki yöntemin başarımlarındaki temel farklılık, konuşmacı ses niteliklerinin farklı şekillerde modellenmesinden kaynaklanmaktadır. VN yönteminde bu modelleme sınırlı sayıdaki kod yöneylerinin belirlenmesi ile olmaktadır. TGM'in üstünlük sağladığı nokta ise aynı modellenmenin belirli sayıda sürekli Gauss yoğunluk işlevleri ile yapılmasıdır.

Kullanılan bu iki yöntemden farklı olarak yeni bir YSA yapısı ile konuşmacı tanıma denemeleri yapılmıştır [6], YSA insan duyma ve algılama sisteminin modellenmesinde araştırmacılara geleneksel istatistiksel yöntemlerden farklı bir yaklaşım sunmaktadır. Bir konuşmacı tanıma sisteminin ilk aşaması olan tanıma ve sına ma yöney öbeklerinin oluşumunda, günümüze kadar kullanılan yöntemlere bir alternatif olarak yeni bir yöntem üzerinde yapılmakta olan çalışmalar devam etmektedir.

Kaynakça

- [1] "Getting started with darpa TIMIT CD-ROM: an acoustic phonetic continuous speech database", National Institute of Standards and Technology (NIST), Gaithersburg, MD.
- [2] Linde Y., Buzo A. ve Gray R.M., "An algorithm for vector quantizer design", IEEE Trans. Comm., Cilt. 20, s. 84-95, 1980.
- [3] Reynolds D. A. ve Rose R.C., "Robust text-independent speaker identification using Gaussian mixture speaker models", IEEE Trans. Speech and Audio Processing, Cilt. 3, s. 72-83, 1995.
- [4] Rosenberg A.E., ve Soong F.K., "Evaluation of a vector quantization talker recognition system in text independent and text dependent modes", Computer Speech and Language, Cilt. 22, s. 143-157, 1987.
- [5] Tishby N. Z., "On the application of mixture AR hidden Markov models to text independent speaker recognition", IEEE Trans. Signal Processing, Cilt. 39, s. 563-570, 1991.
- [6] Orman Ö. D ve Arslan L. M. "A comparative study on closed set speaker identification using RBF Network and Modular Networks", TAINN'2000 de sözlü sunum için kabul edildi.