

Konuřmacı Tanımayı İnceleyen Yeni Bir Dinleyici Sınaması

[†]Özgür Devrim Orman, ^{*}Levent M. Arslan

[†]TÜBİTAK

Ulusal Elektronik ve Kriptoloji Arařtırma Enstitüsü, Gebze, 41470 Kocaeli

oorman@uekae.tubitak.gov.tr

^{*}Boğaziçi Üniversitesi

Elektrik-Elektronik Mühendisliđi Bölümü, Bebek, 80815 İstanbul

arslanle@boun.edu.tr

Öz

İnsanlardaki konuşan kişinin ayırdına varma yetisi günlük hayatın sürdürülmesinde çok önemli bir yer tutmaktadır. Gözönüne alınabilecek pek çok durumda konuşan kişinin kimliğine göre tavır alınması ve verilecek cevabın farklılaşması davranışları düşünüldüğünde bu yetinin varlığının önemi netleşmektedir. Öte yandan, insan konuşmacı tanıma yetisine benzetim yapılarak geliştirilen sistemlerde genel olarak salt insan duyma sistemi ile sınırlı kalındığı görülmektedir. Bu çalışmada geliştirilip uygulanmış olan yeni bir dinleyici sınaması yardımıyla insan duyma sisteminin konuşmacı tanımadaki modeli irdelenip, kabul edilmekte olan yaklaşımla karşılaştırılmaktadır.

1 Giriş

Konuřmacı Tanıma (KT) kavramı sunulan sözden konuşan kişinin kimliğinin çıkarsanmasını kapsar. Günlük hayatta insanlar bu yetiye hergün istemsizce defalarca kez başvurulmaktadır. Örneğın sırtımız dönük olduđu halde seslenen bir tanıdığımızı yüzümüzü dönmeden tanımamız, evimizde kapımızı çalan kişini kimliğini sesinden anlayıp ona göre açıp açmamaya karar vermemiz ve daha pek çok durum aynı yetinin kullanımını içerir. Otomatik Konuşmacı Tanıma (OKT) kuramı da insan duyma sistemi modellenerek geliştirilmiştir. Uygulamada OKT dallara ayrılmaktadır: Konuşmacı Doğrulama (KD) “Speaker Verification” (bir konuşmacının kimliğinin öne sürdüğü kimlikle örtüşüp örtüşmediğinin söz karakteristiklerinin karşılaştırılması yoluyla belirlenimi), Kapalı Küme Konuşmacı Belirlenim (KKKB) “Closed-Set Speaker Identification” (sunulan konuşmacının belirli bir konuşmacı kümesine ait olduđu varsayımına dayanarak sözkonusu küme içindeki konuşmacıların söz karakteristikleri ile sunulan konuşmacının söz karakteristiklerinin karşılaştırılması yoluyla konuşmacının kimliğinin belirlenmesi), Açık Küme Konuşmacı Belirlenim (AKKB) “Open-Set Speaker Identification” (KKKB’den farklı olarak sunulan konuşmacının küme dışından olabileceği olasılığı kabul edilerek küme içindeki konuşmacıların söz karakteristikleri ile sunulan konuşmacının söz karakteristiklerinin karşılaştırılması yoluyla konuşmacının kimliğinin ve küme içinden olup olmadığının belirlenmesi). OKT’nin kullanım alanlarına otomatik giriş kontrolü, adli verilerin işlenmesi ve suçluların izlenmesi örnek olarak verilebilir. Yapılan OKT sistemi gerçeklemederinde insan duyma sistemi yapıla gelmiş psikoakustik sınamalarla belirlenmiş [9] doğrusal olmayan bir sıklık skalasına göre dizilmiş süzgeç öbeği ile modellenmektedir [4]. Bu skalaya göre (mel skalası olarak adlandırılmaktadır) 1 kHz’e kadar olan sıklık değerleri doğrusal ötesindeki sıklık değerleri ise logaritmik bir izdüşüm ile yeni bir sıklık düzlemine dönüştürülmektedir. Salt bir sıklık dönüşümü içeren yaklaşım duyma sisteminin ötesinde yapılan ardıl işlemlerin etkisini gözardı ediyor görünmektedir. Yapmış olduğumuz çalışmayla geliştirilmiş olan dinleyici sınamasında band geçiren süzgeçlenerek sıklıkta kısıtlanmış sözler kullanılarak farklı sıklık bandlarında konuşmacıların tanınabilirliği değerlendirilmektedir. Bu yolla OKT’de kullanıla gelen frekans skalasının öne sürüldüğü şekilde insan duyma sistemi modelinin KT’deki davranışına uygun olup olmadığı irdelenmektedir.

Önceki çalışmalarda geliştirilmiş konuşmacının ayırd edilmesi temeline dayanan dinleyici sınamalarında KB [3], KT [1,2] ve Ses Kodlayıcılarında Konuşmacının Tanınabilirliğinin Korunumu (SKKTK) [5,6,7,8] incelenmeye çalışılmıştır. KB başarımını incelemek için yapılan dinleyici sınamalarında konuşmacının dinleyiciler tarafından önceden tanınıyor olmasının sonuçlar üzerinde belirleyici etkisi olduğu,

konuşmacıların önceden tanınmaması durumunda ise söz uzunluğu ve eğitim zamanı ile sunum zamanı arasındaki sürenin önemli olduğu belirtilmiştir [3]. Aynı sonuca, Konuşmacı Tanıma için yapılmış dinleyici sınamalarında da varılmış ve dinleyicinin konuşmacıyı önceden tanıyor olduğu durumlardaki değerlendirme sonuçlarında bu ön bilgi birikimini (o konuşmacıya ait çeşitli durumlardaki değişik konuşma anıları) yansıtmamasının kaçınılmaz olduğu vurgulanmıştır [2]. SKKTK için yapılmış çalışmaların bir kısmındaysa önceden tanınmayan konuşmacılar [6] ve önceden tanınan konuşmacılar [7] için ayrı sınamalar geliştirilerek ilintinin başarımlar üzerindeki etkisini yalıtma yolunda çaba gösterilmiştir. Özde, bütün bu çalışmalardan elde edilebilecek ortak vargılar; konuşan kişinin önceden tanınmasının başarımlar üzerinde çok belirleyici olması ve hepsinden önemlisi yapılan sorgulama yönteminin ve izlenen niteliklerin dinleyici sınamaları için sözkonusu ilintinin yalıtılmasını güvence altına alamamasının yanında, konuşmacının önceden tanınmadığı durumlarda da çok öznel bir nitelik olan dinleyici belleğinin sonucun üzerinde belirleyici olmasıdır. Bu çalışmada izlenen yaklaşım, sözü edilen çalışmalardan farklı olarak dinleyicilerin kişinin tanınabilirliğini kısa aralıklarla verilen asıl ve band geçiren süzgeçlenmiş söz çiftlerini karşılaştırarak değerlendirmesidir. Söz çiftlerinin karşılaştırmasına dayalı yaklaşımın ardında yatan düşünce bu yolla hem dinleyici bellek performanslarından kaynaklanabilecek olası farklılıkların azaltılması hem de konuşmacının önceden tanınıyor olmasının önemsizleştirilmesidir.

Yapılan çalışmanın sunumu dört bölüme ayrılmıştır. Konuşmacı Tanıma Dinleyici Sınaması (KTDS) kuramı ve kullanılmış olduğumuz istatistiksel yöntem ikinci bölümde açıklanmaktadır. Bu çalışmada geliştirilen KTDS üçüncü bölümde söz kayıtlarının toplanması ve sınamanın uygulanmasının detaylarını içerecek şekilde anlatılmaktadır, aynı bölümde dinleyici değerlendirmelerini istatistiksel çözümlenmesi de verilmektedir. Dördüncü bölüm çalışma sonucu ortaya çıkan vargıları içermektedir.

2 Konuşmacı Tanıma Dinleyici Sınaması Kuramı

Konunun özneliği ve incelenen sistemin karmaşıklığı gözönüne alındığında günümüze kadar yapılan çalışmalarda daha önce de değinmiş olduğumuz temel birtakım noktalara varılmış olmakla birlikte insan duyma sisteminin gerçekleştirdiği işlemler ve ötesindeki ardışıl süreçlerin incelenmesi çeşitli disiplinlerdeki yeni bulgular ile gelişerek sürmektedir. Bu bölümde KT üzerine yapılmış önemli referans çalışmalarda belirtilen dinleyici sınaması uygulamalarını göz önüne alarak konunun kuramsal temeli üzerinde durulacak ve dinleyici değerlendirmelerinin istatistiksel olarak doğruluğunun sorgulandığı Değişinti Çözümlemesi “Analysis of Variance” (DÇ) yöntemi anlatılacaktır.

Atal’ın çalışmasında [1] dört farklı KTDS uygulaması ele alınmıştır. Verilen ilk uygulamada dinleyiciler önceden tanışık olmadıkları beş konuşmacının okudukları bir paragrafta göre karar vermektedirler. Bu uygulamada eğitim ve sınama arasında geçen süreyle konuşmacı tanıma başarımlarının ilintili olduğu gözlenmiştir. İkinci uygulamadaysa konuşmacıların tanınabilirliğiyle sunulan sözün süresi ve fonetik içeriğinin ilintisi ele alınmıştır. Deney kümesi on konuşmacı ve konuşmacıları önceden tanıyan on altı dinleyiciden oluşmaktadır. Dinleyicilerin değerlendirmelerinde oluşabilecek olası bir koşullanmayı engellemek için deney sırasında farklı konuşmacıların söz kayıtlarına erişimi kısıtlanarak sınama yapılmıştır. Bu sınamadan elde edilen sonuçlara göre sunulan sözün içerdiği sesçik “phoneme” miktarı arttığında KT başarımlarında da artış görülmektedir, aynı sonuçlarda dinleyici değerlendirmeleri de farklılıklar göstermektedir. Üçüncü uygulamadaysa on dinleyici ve dinleyicilerin önceden tanışık oldukları sekiz konuşmacı yer almaktadır, ikinci uygulamadan farklı olarak burada dinleyicilerin sınama esnasında diğer konuşmacıların kayıtlarına erişimi serbest bırakılmıştır. Gözlenen KT başarımları beklenildiği gibi kullanılan sözcük, konuşmacı ve dinleyiciye bağlı olarak farklılık göstermektedir. Öte yandan sonuçlarda sunulan sözün içerdiği hece sayısının ikiden fazla olmasının büyük bir artışa sebep olmadığı da gözlenmiştir. Dördüncü ve son uygulamaysa KB’ye yönelik bir çalışmadır, burada dinleyiciler kendilerine sunulan söz çiftlerindeki konuşmacıların aynı ya da farklı kişiler olduklarına karar vermektedirler. Elde edilen sonuçlar karşılaştırılan Otomatik Konuşmacı Belirleme sistemlerinin başarımlarıyla aynı bulunmuştur.

KT alanında diğer bir önemli referans çalışma ise O’Shaughnessy’ye aittir [2]. Bu çalışmada dinleyicilerin tanışık oldukları konuşmacıları tanıma başarımlarının yüksek olduğu ve dinleyicilere karar vermeleri için sunulacak söz uzunluğunun 2–3 saniye olmasının yeterliliği açıklanmıştır. O’Shaughnessy’nin vermiş olduğu örnek uygulamalardan en dikkat çekici olan kırk beş tanınmış kişiyi içeren bir konuşmacı kümesiyle

yapılan sınamada sözler ters yönde sunulularak fonetik içeriğin başarım üzerindeki etkisinin dışlanması sağlanmıştır.

Sözü edilen uygulamalardan görüldüğü gibi probleme farklı açılardan yaklaşılabilir, insan duyma sisteminin KT başarımını'nın tanışıklıkla, sunulan sözün uzunluğuyla ya da söz kaydının dinleyiciye tersten alınmasıyla etkisi azaltılmaya çalışılan içerik bilgisiyle olan ilintisi değişik yaklaşımlarla ele alınmıştır. Özde aynı, uygulamadaysa farklılık gösteren bir dinleyici sınaması türü de SKKTK'yi ele almaktadır. Kodlanmış söz işaretinin asıl söz işaretinden farklılıklar göstermesi nedeniyle söz kodlayıcıların konuşan kişinin kimliğini dinleyiciye doğru iletip iletmediğinin dinleyici sınamalarıyla incelenmesi gerekmektedir. Yapılmış bulunan çalışmalarda varılan sonuçlara göre dinleyicilerle tanışık olmayan bir konuşmacı kümesi kullanılması halinde de dinleyici belleği ve az sayıda konuşmacı içeren sınamalarda konuşmacı kümesinin oluşturulması da sonuç üzerinde etkilidir [6], konuşmacı ile dinleyici tanışıklığının sonucu etkilemesinin yanında dinleyiciler ve konuşmacılar arasındaki tanışıklık ölçüsü kişiden kişiye değişken olacağından sınamaların ele alınmasında bir sorun oluşturmaktadır [7], KT başarımında sözdeki yüksek sıklık değerlerinin band genişliğinin büyük olmasının önemli olduğu da gözlenmiştir [8].

Önceki çalışmalardan görüleceği gibi dinleyici tanışıklığı KT başarımını belirleyen önemli bir etmendir, genel olarak dinleyici sınamalarının tümünde dinleyiciye sunulan sorgulanma yöntemi sonucu oldukça değiştirmektedir. Bu çalışmada tercih edilen yöntem asıl söz ile farklı sıklık bandlarındaki band geçiren süzgeçlerden oluşan süzgeç öbeğindeki bir süzgeçten geçirilmiş olan sözü söyleyen kişilerin kimliklerinin benzer olma miktarlarının dinleyiciler tarafından değerlendirilmesidir. Söz çiftlerinin sunumunda asıl söz ile süzgeçlenmiş söz arasındaki zaman dilimi kısa tutularak dinleyici belleğinden kaynaklanacak olası etkiler kısıtlanmaya çalışılmıştır. Sunum için kullanılacak sözcükler de konuşmacılar ve dinleyicilerde en az duygusal uyarıya sebep olacak şekilde ve fonetik dağılımın dengeliğine dikkat edilerek seçilmiştir.

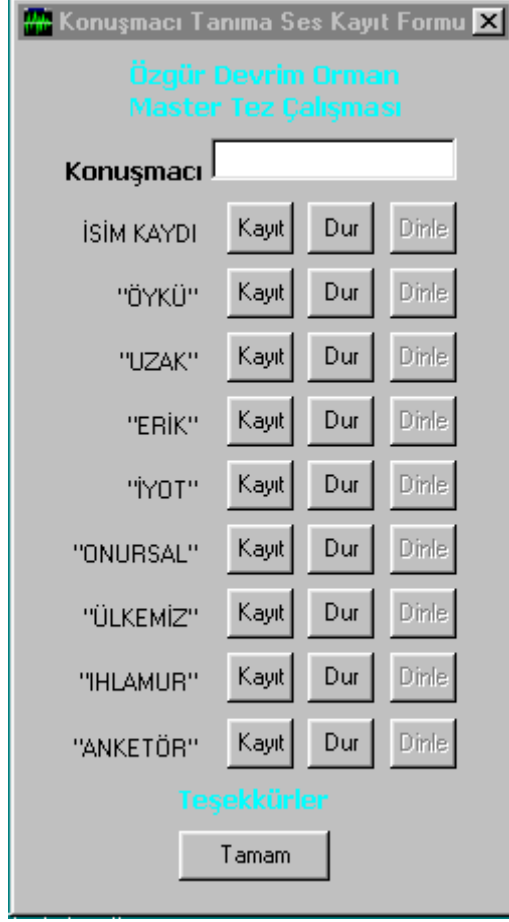
Yapılan dinleyici değerlendirmelerinin istatistiksel doğruluğunu gösteren GÇ yöntemi ele aldığımız farklı sıklık bandlarındaki konuşmacı tanınabilirlik değerlendirmelerinde bandlar arasında değerlendirme farklılıklarının olup olmadığını vermektedir. Kullanılan sıklık bandı sayısını k'ile ve her bir band için elde ettiğimiz değerlendirme sayısını da n'ile gösterirsek, sonuçta doğruluk kararı vermek için kullanacağımız F-istatistiğini aşağıdaki şekilde elde edebiliriz. Denklem 1'de i. sıklık bandındaki j. değerlendirme sonucu y_{ij} ' ile gösterilmiştir.

$$F = \frac{(\sum_{i=1}^k (\sum_{j=1}^n y_{ij})^2 / n - (\sum_{i=1}^k (\sum_{j=1}^n y_{ij} / n) / k)^2 / kn) / (k-1)}{(\sum_{i=1}^k \sum_{j=1}^n y_{ij}^2 - \sum_{i=1}^k (\sum_{j=1}^n y_{ij})^2 / n) / [k \cdot (n-1)]} \quad (1)$$

3 Önerilen Konuşmacı Tanıma Dinleyici Sınaması Uygulaması

Yapılan çalışma ardışık dört evreye bölünerek açıklanabilir; konuşmacılardan ses kayıtlarının alınması, bu kayıtların dinleyiciye sunulacak yapıya getirilmesi, dinleyicilerin değerlendirme yapımları ve yapılan değerlendirmelerin istatistiksel incelemesi.

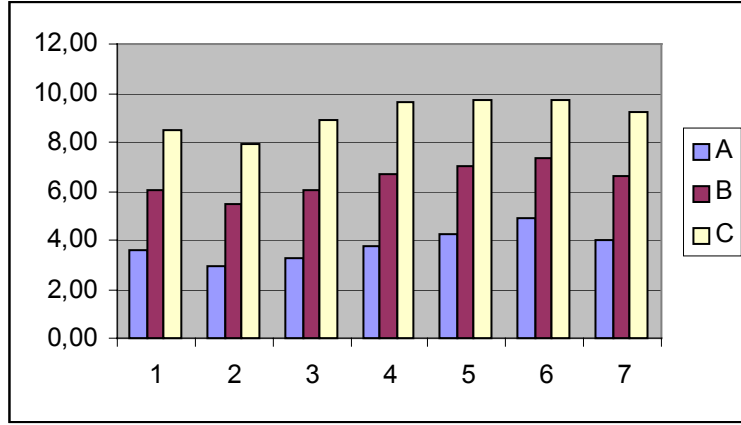
İlk evrede iki bayan ve sekiz erkek konuşmacıdan oluşan bir konuşmacı kümesinden gürültüsüz ortamda ses kayıtları alınmıştır. Konuşmacılara kayıt öncesinde kişisel detayları ile sınamaya için çok önemli olan seslerini etkileyebilecek olası hastalık durumlarını ifade eden bir kayıt belgesi doldurtulmuştur. Ses Kayıt Formu olarak adlandırılan bu belge aynı zamanda konuşmacıların kayıta okuyacakları sözcükleri de içermektedir. Her konuşmacıdan kullanıcıya arkadaş olarak hazırlanmış bir yazılım kullanarak önce isimlerini sonra da akustik dengelik gözönüne alınarak belirlenmiş sekiz sözcüğü ("öykü", "uzak", "erik", "iyot", "onursal", "ülkemiz", "ihlamur", "anketör") kaydetmeleri istenmiş ve tüm kayıtlar aynı bilgisayar üzerinde 16kHz, 16 bit doğrusal veri yapısında gerçekleştirilmiştir, söz konusu programın bir ekran görüntüsü Şekil 1'de verilmiştir.



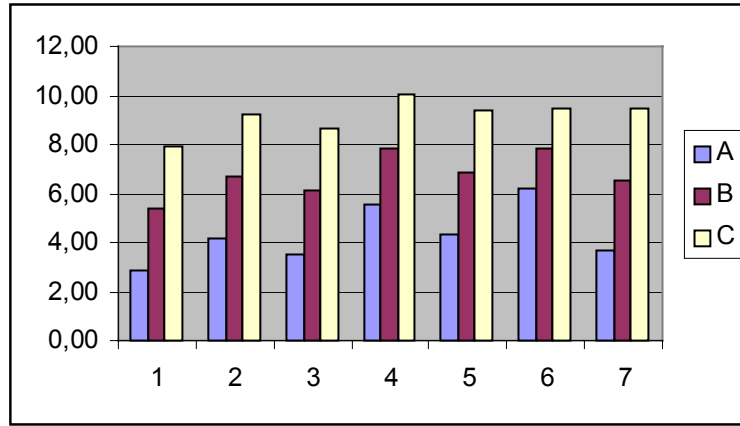
Şekil 1. PC üzerindeki söz kayıt arayüzü.

İkinci evrede, kaydedilmiş sözcükler merkez sıklık değerleri 1kHz den 7kHz'e kadar olan 1kHz band genişliğindeki yedi ayrı band geçiren süzgeçten geçirilmiştir. Elde edilmiş süzgeçlenmiş kayıtlar üzerinde öncelikle eğitim kümesi ve sınama kümeleri ayrışımına gidilmiş, bir erkek ve bir bayan konuşmacının isimlerinin asıllarıyla tüm filtrelenmiş durumları eğitim için ayrılmıştır. Dinleyicilere sunmak için sekiz farklı sınama kümesi oluşturulmuş ve her bir sınama kümesinde sırasıyla 17-18-18-17 söz çifti içeren dört ayrı oturuma bölünmüştür. Sınama kümesi oturumlara bölünerek dinleyicilere oturumlar arasında dinlenme olanağı verilmiştir. Ayrıca söz çiftini oluşturan asıl ve süzgeçlenmiş sözlerin sunumu arasında iki saniye, söz çiftlerinin sunumu arasında da beş saniyelik boşluk bırakılmıştır. Beş saniyelik boşlukta dinleyicilerden değerlendirmelerini kayda geçirmeleri istenmiştir. Eğitim ve sınama kümeleri DATLink kullanılarak DAT kasetlere aktarılıp, taşınabilir bir DAT çaların kullanımı olanaklı kılınmış, böylece farklı yaş ve iş guruplarındaki dinleyicilerin sınamaya katılmaları mümkün olmuştur.

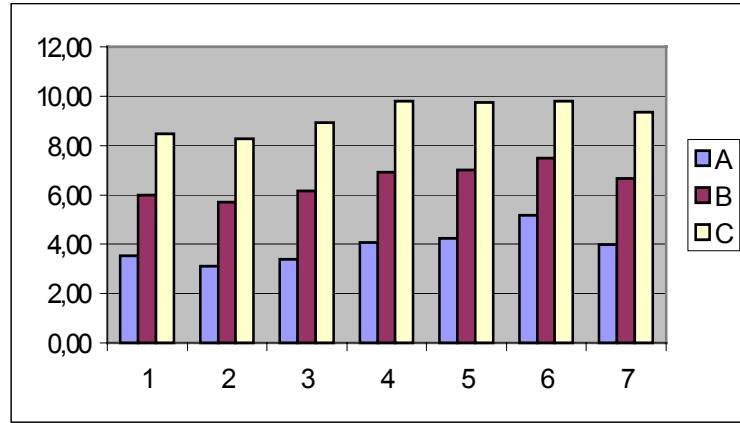
Üçüncü evre olan dinleyici değerlendirmelerinde, dinleyicilerin ortamın etkisinden yalıtımının sağlamak için kayıtlar kulaklık kullanılarak dinletilmiştir. Dinleyici sınamaya başladığında önce eğitim kümesini dinleyip sınama içinde karşılaşacağı değişiklikler hakkında alışkanlık kazandıktan sonra, dinleyicinin izniyle sınama bölümüne geçilmiştir. Dinleyiciler her sözcük çifti için süzgeçlenmiş sözde aslını söyleyen kişinin ne kadar tanınabilir olduğunu sorgulamış ve değerlendirmelerini en kötü için bir ve en iyi için on diye tanımlanmış skala üzerindeki tamsayıları karşı düşürerek yapmışlardır. Dördüncü evre olan istatistiksel incelemede bayan, erkek ve tüm konuşmacılar için dinleyicilerin yapmış oldukları değerlendirmelerin ortalama ve değişimleri belirlenmiştir. Elde edilmiş sonuçlar Şekil 2,3 ve 4'te gösterilmektedir. Şekillerde B ortalama değeri, A ve C'ise sırasıyla aynı bandtaki standart sapma değeri ortalamadan çıkarılıp eklenerek bulunmuş değerleri göstermektedir.



Şekil 2. Salt erkek konuşmacılar için KT başarımları.



Şekil 3. Salt bayan konuşmacılar için KT başarımları.



Şekil 4. Tüm konuşmacılar için KT başarımları.

Elde edilmiş olan dinleyici sınav sonuçlarının ortak karakteristiği 4.,5. ve 6. Sıklık bandlarında KT başarımının yüksek olarak gözlenmiş olmasıdır, sözkonusu bandlar 3 kHz ile 6 kHz sıklık aralığını kapsamaktadır. Varılan bu sonuç önceki bir çalışmada da belirtilmiş olan sunulan sözün içerdiği yüksek frekans band genişliğinin başarım üzerinde etkili olduğu bulgusuyla örtüşmektedir. Sonuçlardaki değişimlerin yüksek olması daha önceki çalışmalarda da gözlenmiş olan dinleyici değerlendirmelerindeki farklılıkların yapılan çalışmada da ortaya çıkmasına karşılık gelmektedir. Dinleyici değerlendirmelerinin istatistiksel doğruluğu DÇ metodu kullanılarak gösterilmiştir. Uygulanan DÇ metodu sonucunda bayan

konuşmacılar için $0.53 \cdot 10^{-2}$, erkek konuşmacılar için $2.23 \cdot 10^{-8}$ ve tüm konuşmacılar için $4.69 \cdot 10^{-11}$ değerleri elde edilmiştir.

4 Vargılar

Yapılan çalışmada özgün bir sına yöntemi önerilmiş olup elde edilen bulgular önceki çalışmalarla örtüşmektedir. KT'yi inceleyen dinleyici sınaalarında dinleyicilerin konuşmacılarla tanışık olmaması durumunda dinleyici belleğinin performansı, birbiriyle tanışık olan konuşmacılar ve dinleyicilerden oluşan sınaalardaysa aralarındaki tanışıklık miktarının kişiden kişiye farklılık göstermesinin değerlendirme sonuçlarında çok etkili olduğu belirtilmiş olup, önerdiğimiz yöntemde bu etkiler en azlanmaya çalışılmıştır. Söz çiftlerinin karşılaştırılması dinleyici belleğine bağımlılığı azaltmaktadır, öte yandan yapılan sorgulamada da dinleyicinin değerlendirmesi kısa aralıkla sunulan temiz ve süzgeçlenmiş sözler üzerinde verilecek bir kararda kısıtlanarak dinleyicinin yargılarından kaynaklanabilecek yanlılığının önüne geçilmeye çalışılmıştır. Elde edilen bulgulara göre 3 kHz–6 kHz sıklık bandının konuşmacı tanıma için önemli olduğu belirlenmiştir. Varılan sonuç, OKT sistemleri için konuşmacı tanıma sürecinin modellenmesinde salt insan duyma sistemi modelinin yeterli olmadığını, duyma sisteminin ötesinde yapılan ardıl işlemlerin sonucunda bilinen lineer olmayan sıklık skalasından farklı bir skala elde edilmekte olduğunu göstermektedir.

5 Teşekkür

Yapılmış olan dinleyici sınamasına konuşmacı ya da dinleyici olarak katılan TÜBİTAK–UEKAE personeline, Dokuz Eylül Üniversitesi ve İzmir Yüksek Teknoloji Enstitüsündeki Araştırma Görevlisi arkadaşlara teşekkürlerimizi sunarız.

Kaynakça

- [1]. Atal, B. S., “Automatic recognition of speakers from their voices”, Proc. IEEE, c. 64, s. 460-474, 1976.
- [2]. Doddington, G. R., “Speaker recognition–identifying people by their voices”, Proc. of IEEE, c. 73, s. 1651–1664, Kasım 1985.
- [3]. O’Shaughnessy, D., “Speaker recognition”, IEEE ASSP Magazine, s. 4–17, 1986.
- [4]. Reynolds D. A. ve Rose R. C., “Robust text-independent speaker identification using Gaussian mixture speaker models”, IEEE Trans. Speech and Audio Processing, c. 3, s. 72–83, 1995.
- [5]. Schmidt–Nielsen A. ve Brock D., “Speaker recognizability testing for voice coders”, Int. Conf. on Acous., Speech, and Sig. Proc., s. 1149–1152, 1996.
- [6]. Schmidt–Nielsen A. ve Stern K. R., “Recognition of previously unfamiliar speakers as a function of narrowband processing and speaker selection”, J. Acoust. Soc. America, c. 79, s. 1174–1177, 1986.
- [7]. Schmidt–Nielsen A. ve Stern K. R., “Identification of known voices as a function of familiarity and narrowband processing”, J. Acoust. Soc. America, c. 77, s. 658–663, 1985.
- [8]. Uzdy Z., “Human speaker recognition performance of LPC voice processors”, IEEE Trans. Acoust. Speech, Signal Processing, c. ASSP-33, s. 752–753, 1985.
- [9]. Zwicker E. ve Fastl H., *Psychoacoustics: Facts and Models*, Springer–Verlag, Berlin, 1990.