mühendislikdergisi
Dicle Üniversitesi Mühendislik Fakültesi
Cilt: 8, Sayı: 1, 77-88
Mart 2017

Mobil Metne Bağımlı Tek Cümle Konuşmacı Tanıma Uygulamasında Kayıttan Sahte Doğrulama

Osman BÜYÜK[1]

[1] *Kocaeli Üniversitesi, Elektronik ve Haberleşme Mühendisliği Bölümü,Umuttepe Yerleşkesi,41380, Kocaeli.*

Makale Gönderme Tarihi: 07.06.2016 Makale Kabul Tarihi: 19.10.2016

Öz

Son yıllarda akıllı telefon gibi mobil araçların kullanımındaki hızlı artış farklı teknolojileri bu platformlar için gerçekleştirmeyi önemli bir sektör haline getirmiştir. Mobil uygulama sayısındaki bu artış bu uygulamalardaki güvenlik meselesini de ön plana çıkarmıştır. Konuşmacının sesinden kimliğinin otomatik olarak belirlenmesini sağlayan konuşmacı tanıma teknolojisi kişisel bilgi güvenliği gerektiren mobil uygulamalarda güvenlik açığını gidermek için kullanılabilir.

Metne bağımlı tek cümle konuşmacı tanıma uygulamasında konuşmacılar eğitim ve tanıma sırasında ortak parola cümlesini tekrar ederler. Eğitim ve tanımada aynı metnin tekrarlaması tanıma performansını arttırdığı gibi kullanım kolaylığı da sağlamaktadır. Bununla birlikte tek cümle uygulamaları özellikle kayıttan sahte doğrulama ataklarına karşı son derece savunmasızdır. Bu çalışmada metne bağımlı tek cümle uygulamasının kayıttan sahte doğrulama ataklarına karşı dayanıklılığı test edilmiştir.

Bu çalışmada mobil araçlar için geliştirilecek tek cümle uygulamasının kayıttan sahte doğrulama ataklarına karşı dayanıklılığını test edebilmek için yeni bir konuşmacı tanıma veri tabanı oluşturulmuştur. Bu veri tabanında 124 konuşmacı (62 bayan + 62 bay) 2 ayrı oturumda belirlenen parola cümlesini tekrar etmiştir. Kayıtlar 2 farklı akıllı telefon kullanılarak alınmıştır. Bu veri tabanı ile kayıttan sahte doğrulama saldırıları simüle edilmiştir.

Gauss karışım modeli (Gaussian mixture models - GKM) metinden bağımsız uygulamalarda en sık kullanılan yöntemlerdendir. Saklı Markov model (hidden Markov model - SMM) tabanlı yöntemler ise metne bağımlı uygulamalarda artikülasyon bilgisinden daha iyi faydalandıkları için tercih edilmektedir. Son dönemlerde kanal uyuşmazlığı problemini gidermek için i-vektör/PLDA yöntemi önerilmiş ve özellikle metinden bağımsız uygulamalarda son derece başarılı sonuçlar vermiştir.

Bu çalışmada GKM, cümle SMM ve i-vektör/PLDA yöntemleri mobil metne bağımlı tek cümle uygulamasında kayıttan sahte doğrulama ataklarına karşı test edilmiştir. Deneylerde tüm yöntemlerin sahte doğrulama saldırılarından önemli ölçüde etkilendiği gözlenmiştir. Yaptığımız testlerde eşit hata oranları normal sahte doğrulama denemelerinde %0.5-1 aralığındayken, kayıttan sahte doğrulama denemeleriyle %10-25 aralığına yükselmiştir.

Anahtar Kelimeler: *Konuşmacı tanıma; metne bağımlı tek cümle; kayıttan sahte doğrulama;mobil araçlar; akıllı telefonlar.*

*Yazışmaların yapılacağı yazar: Osman BÜYÜK. osman.buyuk@*kocaeli.edu.tr*; Tel: +90 (262) 3033389

Giriş

Son yıllarda akıllı telefon, tablet gibi mobil araçların kullanımı hızlı bir şekilde artmıştır. 2014 yılında yapılan bir araştırmaya göre tüm dünyada kullanılan mobil araç sayısı 2013 yılının sonunda dünya nüfusunu geçmiştir. Aynı araştırma mobil araç kullanıcılarının 30 günlük süre içerisinde ortalama 6.5 uygulamayı aktif olarak kullandığını ortaya koymuştur (Super Monitoring, 2013). Android market istatistiklerine göre toplam 500 milyon indirme sayısına ulaşmış popüler uygulamalar bulunmaktadır. Mobil uygulama ve kullanıcı sayısındaki bu hızlı artış, bu uygulamalardaki güvenlik durumunu da ön plana çıkarmıştır. Kullanıcıların kimliğinin sesinden belirlenmesini sağlayan konuşmacı tanıma teknolojisi mobil güvenlik konusunda önemli bir alternatiftir.

Konuşmacı tanıma uygulamaları genel olarak iki kategoriye ayrılabilir; metinden bağımsız (text-independent) ve metne bağımlı (text-dependent). Metne bağımlı uygulamalarda kullanıcı önceden belirlenmiş bir metni tekrar eder. Metinden bağımsız uygulamalarda böyle bir metin kısıtlaması yoktur. Gauss karışım modeli (Gaussian mixture model - GKM) özellikle metinden bağımsız uygulamalarda en sık kullanılan yöntemlerden birisidir (Reynolds vd., 2000). Metne bağımlı uygulamalarda ise saklı Markov model (hidden Markov model - SMM) tabanlı yöntemler artikülasyon bilgisinden faydalandıkları için tercih edilmektedir. Son zamanlarda GKM yönteminin üstüne kanal uyuşmazlığı problemini en aza indirgemek için i-vektör/PLDA yöntemi önerilmiş ve özellikle metinden bağımsız uygulamalarda son derece başarılı sonuçlar vermiştir (Kenny, 2012; Sturim vd., 2011; Hasan vd., 2012; Ferre vd., 2013).

Metne bağımlı tek cümle (text dependent single uttteance - MBTC) Türkiye'de farklı sektörlerde pratik olarak kullanılan bir konuşmacı tanıma uygulamasıdır. Bu uygulamada kullanıcılar önceden belirlenmiş tek bir parola cümlesini eğitim ve tanıma sırasında tekrar eder. Eğitim ve tanımada aynı parolanın tekrar edilmesi bu uygulamalardaki tanıma performansını olumlu etkilemektedir. Fakat aynı sebeple tek cümle uygulaması kayıttan sahte doğrulama saldırılarına karşı son derece açık hale gelmektedir.

Konuşmacı tanıma sistemlerinin farklı tekniklerle yapılacak sahte doğrulama ataklarına karşı güvenilirlikleri bu teknolojinin ticari ürünlerde kullanılacak olgunluğa ulaşmasıyla birlikte daha önemli hale gelmiştir. (Wu vd. 2015a)'da sahte doğrulama konusunda yapılan çalışmaların bir özeti sunulmuştur. Bu çalışmada olası sahte doğrulama saldırıları arasında taklit etme, kayıttan tekrar çalma, konuşma sentezi ve konuşmacı dönüştürme teknolojileri sayılmıştır. Bu saldırıların konuşmacı tanıma sistemleri için önemli bir problem teşkil ettiği belirtilmiştir. (Wu vd. 2012)'de konuşma sentezi ve konuşmacı dönüştürme saldırılarını doğrulama öncesinde tespit etmek için değiştirilmiş grup gecikme özniteliklerinin (modified group delay features - DGGÖ) kullanılması önerilmiştir. Çalışmada, DGGÖ öznitelikleri kullanılarak GKM temelli bir sınıflandırma yapılmaktadır. Yapılan deneylerde DGGÖ temelli yöntem sahte doğrulama saldırılarını belirlemede oldukça başarılı sonuçlar vermiştir.

Son yıllarda sahte doğrulama konusuna ilginin artması farklı kurumların bu konuda çalışmalar yapmasına neden olmuştur. Bu kurumlarda yapılan bağımsız çalışmaların karşılaştırılabilir olmaması ortak bir veri tabanı oluşturulması ihtiyacını doğurmuştur. Bu amaçla metinden bağımsız konuşmacı tanıma uygulamaları için ASVspoof2015 veri tabanı oluşturulmuştur. Farklı kurumların katılımıyla ilk sahte doğrulama yarışması 2015 yılında düzenlenmiştir (Wu vd., 2015b; Alam vd., 2015; Chen vd., 2015; Janicki, 2015).

Kayıttan sahte doğrulama konusunda da önceki

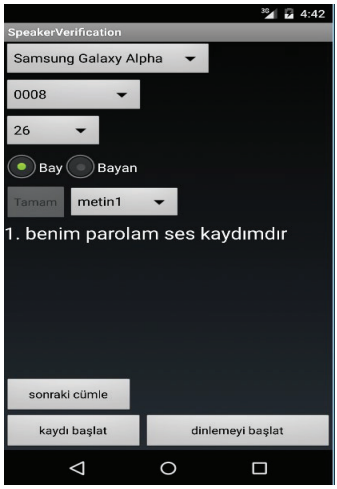
yıllarda yapılmış çalışmalar bulunmaktadır. (Alegre vd., 2014)'te kayıttan sahte doğrulama ataklarının GKM, destek vektör makineleri (support vector machines - DVM), birleşik faktör analizi (joint factor analysis), i-vektör/PLDA gibi farklı konuşmacı tanıma yöntemleri üzerindeki etkisi metinden bağımsız bir veri kullanılarak incelenmiştir. Çalışmada kayıttan sahte doğrulama ataklarının konuşmacı tanıma sistemleri için önemli bir problem teşkil ettiği sonucuna varılmıştır. Ayrıca kayıttan sahte doğrulama ataklarının konuşma sentezi ve konuşmacı dönüştürmeye göre daha kolay elde edilebileceği vurgulanmıştır. (Shang ve Stevenson, 2010)'da kayıttan sahte doğrulama ataklarının etkisi metne bağımlı bir veri tabanında incelenmiştir. Çalışmada, saldırıları tespit etmek için test kayıtlarının sistem tarafından kaydedilmiş önceki denemelerle karşılaştırılması önerilmiştir. (Wu vd., 2014)'te sistem tarafından kaydedilen kayıtların boyutunu küçültmek için spektrogram ikili eşlem (spectrogram bitmap) tabanlı bir yöntem önerilmiştir. Yöntemde spektral tepeler 1, tepe olmayan noktalar ise 0 ile kodlanmaktadır. Deneylerde, önerilen kayıttan sahte doğrulama tespit yönteminin konuşmacı tanıma performansını son derece olumlu etkilediği gözlenmiştir.

Bu çalışmada MBTC uygulamasında GKM, SMM ve i-vektör/PLDA yöntemlerinin kayıttan sahte doğrulama ataklarına karşı dayanıklılığı test edilmiştir. Bu amaçla, mobil uygulamalar için yeni bir konuşmacı tanıma veri tabanı oluşturulmuştur. Veri tabanında 62'si bayan, 62'si bay olmak üzere toplam 124 konuşmacı bulunmaktadır. Kayıtlar 2 ayrı oturumda, 2 farklı akıllı telefon kullanılarak alınmıştır. Veri tabanı kullanılarak kayıttan sahte doğrulama denemeleri simüle edilmiştir. Yaptığımız testlerde üç metodun da kayıttan sahte doğrulama ataklarından ciddi şekilde etkilendiği gözlenmiştir. Testlerde normal sahte doğrulama denemeleriyle %0.5-1 aralığında olan hata oranı, kayıttan sahte doğrulama denemeleriyle %10-20 aralığına çıkmıştır.

Makalenin devamı şu şekilde düzenlenmiştir. Bölüm 2'de mobil konuşmacı tanıma veri tabanının özellikleri anlatılacaktır. Bölüm 3 kullanılan yöntemlere ayrılmıştır. Bölüm 4'te konuşmacı tanıma deney sonuçları paylaşılacaktır. Bölüm 5 tartışma ve sonuca ayrılmıştır.

Mobil Tek Cümle Konuşmacı Tanıma Veri Tabanı

Konuşmacı tanıma veri tabanını oluştururken farklı konuşmacıların ses kayıtlarını doğru ve hızlı bir şekilde alabilmek Şekil 1'de gösterilen arayüz kullanılmıştır. Bu arayüz Android platformu için hazırlanmıştır;



Şekil 1 : Mobil konuşmacı tanıma veri tabanı ses kaydı toplama arayüzü.

Gerçekleştireceğimiz metne bağımlı tek cümle uygulamasında parola cümlesi olarak "benim parolam ses kaydımdır" seçilmiştir. Bu cümlenin seçilmesinde parola anlamı taşımasıyla birlikte Türkçe'deki 8 sesli harften 5'ini içermesi de etkili olmuştur. Belirlenen sesli parola cümlesi 2 farklı Android tabanlı akıllı telefon kullanılarak toplam 124 konuşmacıdan alınmıştır. Konuşmacıların 62'si bay, 62'si bayandır. %60'ı 18-25 yaş arası üniversite öğrencisidir. Veri tabanındaki

konuşmacıların büyük bir kısmının yakın yaş grubunda olması konuşmacı tanıma açısından önemli bir zorluk teşkil etmektedir. Kayıtlar 2 ayrı oturumda alınmıştır. İlk oturuma 124 konuşmacının tamamı, ikinci oturuma 102 konuşmacı katılmıştır. Konuşmacılardan her iki oturumda Şekil 1'deki arayüzü kullanarak parola cümlesini 10 kez tekrar etmiştir. Kayıtlar iki telefondan aynı anda paralel olarak alınmıştır.

Veri tabanındaki tüm kayıtlar gürültülü ofis ortamında alınmıştır. Kayıtlardaki gürültü oranı ofisin yoğunluğuna göre rastgele değişmiştir. Bu durum daha gerçekçi sonuçlar elde etmek için tercih edilmiştir. Bu çalışmada gürültünün konuşma tanıma performansı üzerindeki etkisi incelenmemiştir. Fakat veri tabanındaki kayıtların üzerine farklı oranlarda gürültü eklenerek bu etki de incelenebilir. Veri tabanındaki kayıtların tümü 16 kHz, 16 bit, tek kanal, darbe kod modülasyonu (pulse code modulation - PCM) formatındadır. Topladığımız veri tabanını gelecekte akademik araştırma amacıyla araştırmacıların kullanımına açmayı planlıyoruz.

Yöntem

Bu bölümde GKM, cümle SMM ve i-vektör/PLDA yöntemlerinin ayrıntıları paylaşılacaktır. Her üç yöntemde ortak öznitelik vektörleri kullanılmıştır. Bu özniteliklerin ayrıntıları izleyen alt bölümde verilecektir. Son bölümde kayıttan sahte doğrulama denemelerinin oluşturulmasında kullandığımız yöntem anlatılacaktır.

Öznitelik Çıkarımı

Her üç yöntemde ortak kullanılan öznitelik vektörleri 13 mel-frekans kepstral katsayısından (MFKK) oluşmaktadır. Birinci derece delta katsayılarının da vektöre eklenmesiyle 26 boyutlu öznitelik vektörü elde edilmiştir. MFKK çıkarımı sırasında 25 milisaniye pencere boyutu, 10 milisaniye atlama boyutu kullanılmıştır. Öznitelik vektörlerine kepstral ortalama normalizasyonu uygulanmıştır. Veri tabanındaki tüm kayıtlar öznitelik çıkarımından önce baştaki ve sondaki sessizlik kısımlarını atacak bir ön işlemden geçirilmiştir. Bu amaçla mobil uygulamalar için eğitilmiş üçlü-ses (tri-phone) SMM'leri kullanılmıştır. Üçlü-ses SMM'lerin eğitiminde yaklaşık 10 saatlik bir ses verisi kullanılmıştır. SMM eğitim kayıtları farklı akıllı telefonlardan alınmıştır. Sessizlik kısımlarının kesilmesi için cümleler önce SMM'ler ile ses birimlerine ayrılmıştır. Daha sonra sessizlik hizalanan kısımlar kesilmiştir. Öznitelikler sessizlik kısımları kesilmiş kayıtlardan çıkarılmıştır.

Gauss Karışım Modeli (GKM)

GKM yönteminde her konuşmacı belli sayıda karışıma sahip bir Gauss modeli ile modellenir. Yöntemde konuşmacı modelleri genel arka plan (universal background model - GAM) modelinden adapte edilmektedir (Reynolds vd., 2000). Konuşmacı modeli adaptasyonunda maksimum sonsal olasılık (maximum a posteriori - MSO) ya da maksimum olasılık doğrusal regresyon (maximum likelihood lineer regression - MODR) yöntemleri kullanılmaktadır. Bu yöntemler kısıtlı adaptasyon verisi ile güvenilir modeller elde etmek için tercih edilmektedir.

GKM'de, GAM çok sayıda konuşmacının konuşmacıdan bağımsız genel modelini ifade etmektedir. Bu modelin eğitiminde, birden çok konuşmacının ses verisini içeren geniş bir veri tabanı kullanılmaktadır.

Doğrulama skoru elde edilirken, öznitelik vektörlerinin GKM tarafından üretilme olasılığı Denklem 1'deki şekilde hesaplanır;

$$P(\boldsymbol{o}_t|\Lambda) = \sum_{i=1}^{M} w_i N(\boldsymbol{o}_t, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \qquad (1)$$

Her karışımın $\{w_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}$ parametrelerine sahip bir Gauss dağılımı olduğu varsayılır;

$$N(\boldsymbol{o}_t, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}_i|^{1/2}} * exp\left[-\frac{1}{2}(\boldsymbol{o}_t - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i{}^{-1}(\boldsymbol{o}_t - \boldsymbol{\mu}_i)\right] \quad (2)$$

Denklem 1 ve Denklem 2'de *M* karışım sayısını, w_i *i*'inci karışımın ağırlığını, $\boldsymbol{\mu}_i$ *i*'inci karışımın ortalama vektörünü, $\boldsymbol{\Sigma}_i$ *i*'inci karışımın kovaryans matrisini ve $\boldsymbol{o}_t$ *t*'inci öznitelik vektörünü ifade etmektedir.

Öznitelik vektörlerinin birbirinden bağımsız olduğu kabul edilirse, konuşmacı skorunun logaritması Denklem 3'deki gibi hesaplanabilir;

$$\Lambda_S(\boldsymbol{O}) = \sum_{t=1}^{T} log\, P(\boldsymbol{o}_t|\lambda_S) \quad (3)$$

Denklem 3'de gösterildiği şekilde elde edilen konuşmacı skoru GAM skoruyla normalize edilerek, log-olasılık oranı (log-likelihood ratio) elde edilmektedir. Bu oran doğrulama kararının verilmesinde kullanılmaktadır.

Bu çalışmada GAM çok sayıda konuşmacının parola cümlesi kayıtlarından eğitilmiştir. Konuşmacıdan bağımsız GAM 256 karışımlıdır. Konuşmacı modelleri GAM'dan MLLR-MAP tekniği ile adapte edilmiştir (Blouet vd., 2004). Adaptasyon sırasında sadece ortalama değerleri adapte edilirken, karışım ağırlıkları ve kovaryans GAM'dan kopyalanmıştır. Bu yöntemin tercih edilmesinde adaptasyon verisinin kısıtlı olması etkili olmuştur. Doğrulama kararı log-olasılık oranı kullanılarak verilmektedir.

Cümle Saklı Markov Modeli (Cümle SMM)

SMM'ler özellikle konuşma tanıma uygulamalarında en sık kullanılan yöntemlerdendir. SMM yönteminde gözlenen öznitelikler saklı durumlarla (states) hizalanır. Her durum bir GKM ile modellenir. Durumların birbirine geçişleri durum geçiş olasılıkları (state transition probabilities) ile ifade edilir.

GKM özellikle metinden bağımsız uygulamalarda en sık kullanılan yöntem olmasına rağmen, SMM tabanlı yöntemler metne bağımlı uygulamalarda daha çok tercih edilmektedir. Bunun en önemli nedeni SMM'lerin artikülasyon bilgisinden daha iyi faydalanmasıdır. Metne bağımlı tek cümle uygulamasında, eğitim ve doğrulama sırasında aynı cümlenin tekrar edilmesi, sesli parola için tek bir cümle SMM'i eğitilmesine olanak vermektedir. Önceki çalışmalarımızda tek cümle uygulamasında cümle SMM'in, tek-ses SMM (monopohone HMM) ve GKM yöntemlerine göre daha iyi sonuçlar verdiği gözlenmiştir (Buyuk, 2011; Buyuk ve Arslan, 2012).

Cümle SMM yönteminde, parola cümlesi için konuşmacıdan bağımsız bir cümle SMM'i oluşturulmuştur. Konuşmacıdan bağımsız SMM GAM eğitimindeki veri ile eğitilmiştir. Cümle SMM'de 64 durum bulunmaktadır. Durum sayısı parola cümlesindeki ses birimi sayısına orantılı seçilmiştir. Her durum 4 karışım ile modellenmiştir. Böylece, cümle SMM ve GKM yöntemlerinde eşit model büyüklükleri elde edilmiştir. SMM yapısı olarak soldan-sağa durum atlamasız (left-to-right without skip state) yapı tercih edilmiştir. Konuşmacı modelleri, konuşmacıdan bağımsız SMM'den MSO yöntemi kullanılarak adapte edilmiştir. Adaptasyonda sadece ortalama vektörleri uyarlanırken, modeldeki diğer parametreler konuşmacıdan bağımsız modelden kopyalanmıştır. Doğrulama kararı için zorla hizalama olasılıkları (forced alignment likelihoods) kullanılmıştır. Bu olasılıklar konuşmacıdan bağımsız model skoruyla normalize edilmiştir.

i-vektör/PLDA

i-vektör yönteminde öznitelikler az boyutlu toplam değişinti uzayına (total variability space) izdüşürülür. Toplam değişinti uzayı birleşik faktör analizinden farklı olarak hem konuşmacı hem de kanal değişkenliğini içermektedir (Dehak vd., 2011). İzdüşümü gerçekleştirmek için, GAM'ın her karışımı için birinci derece

Baum-Welch istatistikleri toplanır. İstatistikler ardarda eklenerek süper vektör oluşturulur. Bu süper vektörün Denklem 4'teki faktör analizi modeline uyduğu varsayılır (Garcia-Romero ve Espy-Wilson, 2011);

$$M = m + Tw \quad (4)$$

Denklemde m GAM süper vektörü, T düşük dereceli (low rank) dikdörtgen bir matris, w standart normal dağılıma sahip rastgele bir vektördür. Denklemdeki T toplam değişinti matrisi olarak adlandırılmaktadır ve geniş bir veriden beklenti maksimizasyonu (expectation-maximization) algoritması ile eğitilmektedir. i-vektör w'nın nokta MSO kestirimidir (Garcia-Romero ve Espy-Wilson, 2011).

i-vektör/PLDA yönteminde, i-vektörlerin boyutu doğrusal ayırıcı analizi (linear discriminant analysis - DAA) ile biraz daha azaltılmaktadır. i-vektörler elde edildikten sonra, bu vektörlerin olasılıksal üretimsel bir modelden geldiği varsayılarak olasılıksal doğrusal ayırıcı analizi (probabilistic linear discriminant analysis - PLDA) modellemesi uygulanmaktadır. PLDA başlangıçta yüz tanıma için önerilmiş (Prince ve Elder, 2007), daha sonra konuşmacı tanımaya da başarılı bir şekilde uygulanmıştır (Kenny, 2010). PLDA'de her bir i-vektör Denklem 5'teki gibi parçalarına ayrılır;

$$\eta = \mu + \Phi y + \varepsilon \quad (5)$$

Denklemde η i-vektör, μ i-vektörlerin ortalaması, y standart normal dağılıma sahip gizli birim vektördür. ε artık terim olarak adlandırılır ve diğer değişkenler tarafından modellenmeyen değişkenlikleri modeller. Bu terimin sıfır ortalama ve tam kovaryansa sahip bir Gauss dağılımı olduğu varsayılır. PLDA yönteminin parametreleri geniş bir geliştirme verisinden beklenti maksimizasyonu algoritması ile eğitilmektedir. Eğitimden önce genelde i-vektörlerin ortalaması ve uzunluğu normalize edilmekte ve i-vektörler beyazlaştırılmaktadır (Garcia-Romero ve Espy-Wilson, 2011; Sadjadi vd., 2013).

Doğrulama skoru için eğitim, ηe ve test, ηa i-vektörleri arasında Denklem 6'daki hipotez testi uygulanmaktadır;

H1: İki i-vektör aynı gizli vektör tarafından üretilmiştir.

H2: İki i-vektör farklı vektörler tarafından üretilmiştir.

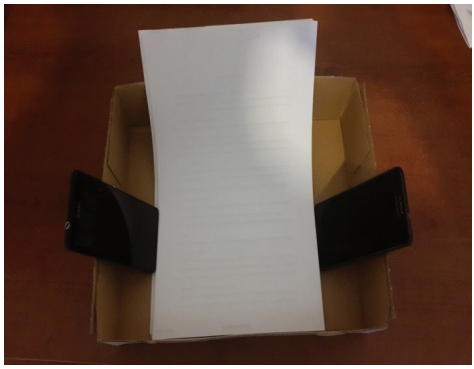
$$score = log\ \frac{p(\eta_e, \eta_a/H_1)}{p(\eta_e/H_2)p(\eta_a/H_2)} \quad (6)$$

Denklem 5 Gauss varsayımında kapalı bir çözüme sahiptir. Skorlama konusunda daha fazla detay için (Garcia-Romero ve Espy-Wilson, 2011; Prince ve Elder, 2007) referanslarına başvurulabilir.

i-vektör/PLDA yönteminin gerçekleştirilmesinde geliştirme verisi olarak GKM yöntemindeki GAM eğitim verisi kullanılmıştır. Daha önce belirtildiği gibi cümle SMM yönteminde konuşmacıdan bağımsız model de aynı veri ile eğitilmiştir. Toplam değişinti matrisinde 150 faktör vardır. Faktör sayısı kısa testler yapılarak belirlenmiştir. Boyut doğrusal ayırıcı analizi ile 75'e indirilmiştir. Elde edilen i-vektörlere PLDA uygulanmıştır.

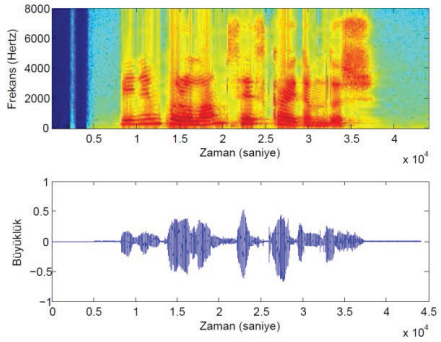
Kayıttan sahte doğrulama

Kayıttan sahte doğrulama saldırılarını simüle etmek için Şekil 2'de gösterilen düzenek kurulmuştur. Bu deney düzeneği kullanılarak konuşmacı tanıma veri tabanındaki tüm kayıtlar bir telefondan çalınıp diğer telefona kaydedilmiştir. Bu kayıtlar kayıttan sahte doğrulama testlerinde kaydedilen telefon hesabına sahte doğrulama denemesi olarak kullanılmıştır;

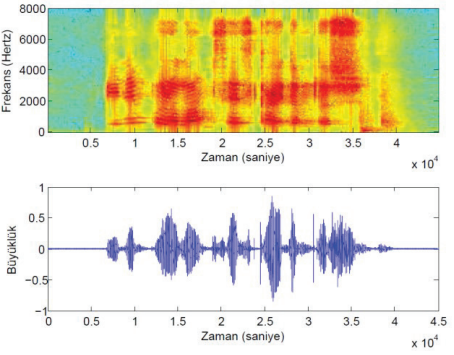


Şekil 2: Kayıttan sahte doğrulama atakları simülasyonu deney ortamı.

Şekil 3 ve Şekil 4'te kayıttan sahte doğrulama deneyleri için kullandığımız kayıtların spektrogram görüntüleri verilmiştir. Şekil 3'de orijinal kayıt, Şekil 4'te aynı kaydın bir telefondan çalınıp diğer telefona kaydedilmiş hali gösterilmektedir. Şekillerde görüldüğü gibi tekrar çalınıp kaydedilen kaydın genel spektrum yapısı orijinal kayda benzemektedir. Fakat, tekrar çalınan kaydın spektrumu ayrıntılı bir şekilde incelendiğinde, özellikle düşük frekanslardaki formantların belirginliğini kaybettiği gözlenmektedir. Bu durum Şekil 2'de gösterilen deney düzeneğinden kaynaklanıyor olabilir. Şekil 2'de görüldüğü gibi kayıttan sahte doğrulama kayıtlarının elde edilmesi sırasında her iki telefon karşılıklı olarak yerleştirilmiş, orijinal kayıtlar otomatik olarak birisinden çalınıp diğerine kaydedilmiştir. Kullanılan telefonlarda mikrofonun yerinin aşağıda ses çıkışının ise yukarıda olması tekrar çalma kayıtlarının kalitesini etkilemiş olabilir. Ayrıca deney düzeneğinde görüldüğü gibi mikrofonun bulunduğu kısım yerle teması nedeniyle büyük oranda kapalı durumda bulunmaktadır. Deney düzeneğinin telefonların ses giriş ve çıkış kısımlarını daha yakın yerleştirilecek şekilde düzenlenmesi orijinale daha yakın kayıtlar elde edilmesini sağlayabilir. Gelecek çalışmalarımızda deney düzeneğinin bu şekilde düzenlenmesi üzerinde durulacaktır. Ayrıca, şekillerde görüldüğü gibi sahte doğrulama kaydındaki geri plan gürültüsü orijinal kayda göre daha yüksektir.



Şekil 3: Mobil konuşmacı veri tabanından örnek bir kayıt.



Şekil 4: Örnek kaydın kayıttan sahte doğrulama deneyleri için bir telefondan çalınıp diğer telefona kaydedilmiş hali.

Deneyler

Konuşmacı tanıma deneylerinde deneme sayısını arttırmak için veri tabanındaki konuşmacılar 6 gruba ayrılmıştır. İlk 5 grupta 20, son grupta 24 test konuşmacısı vardır. Veri tabanındaki kişiler test konuşmacısı olarak sadece bir kez kullanılmıştır. Test grubundaki konuşmacıların cinsiyet dağılımları eşittir.

Her konuşmacının ilk oturumdaki üç kaydı konuşmacı modeli adaptasyonu için ayrılmıştır. Diğer kayıtlar testlerde kullanılmıştır. Test grubunda bulunan konuşmacıların test için

ayrılan kayıtları kendi hesapları için gerçek (target trial), test grubundaki diğer konuşmacıların hesapları için sahte (imposter trial) deneme olarak kullanılmıştır.

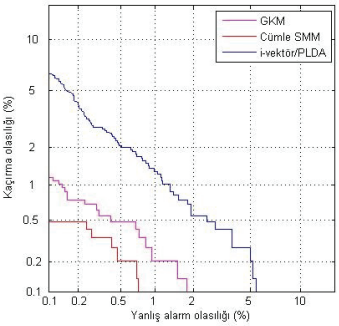
Test konuşmacıları arasında bulunmayan 100 konuşmacı model eğitimi için geliştirme konuşmacısı olarak ayrılmıştır. Her gruptaki geliştirme konuşmacılarının 50'si bayan, 50'si baydır. GKM yönteminde GAM, cümle SMM yönteminde konuşmacıdan bağımsız SMM ve i-vektör/PLDA yönteminde toplam değişinti matrisi eğitimi geliştirme konuşmacılarının tüm kayıtları kullanılarak yapılmıştır.

Deneyler sırasında GKM, cümle SMM ve i-vektör/PLDA yöntemleri için sırasıyla Becars (Blouet vd., 2004), HTK (Young vd., 2006) ve MSR (Sadjadi vd., 2013) kütüphaneleri kullanılmıştır. Bu kütüphaneler daha önceki çalışmalarda bu yöntemler için başarılı sonuçlar verdiği için tercih edilmiştir.

Tablo 1'de elde ettiğimiz konuşmacı tanıma sonuçları verilmiştir. Tablodaki eşit hata oranları (equal error rate - EHO) 6 test grubundaki gerçek ve sahte denemeler birleştirilerek hesaplanmıştır. Bu deneyde toplam 1462 gerçek ve 28894 sahte doğrulama denemesi yapılmıştır. Şekil 5'te yöntemlerin tanıma hata oranı eğrileri (detection error trade-off curve) gösterilmektedir.

Tablo 1: Normal sahte doğrulama deneyi yüzde eşit hata oranları (%EHO).

	%EHO
GKM	0.478
CÜMLE SMM	0.342
İ-VEKTÖR/PLDA	1.094



Şekil 5: Normal sahte doğrulama deneyi tanıma hata oranı eğrisi.

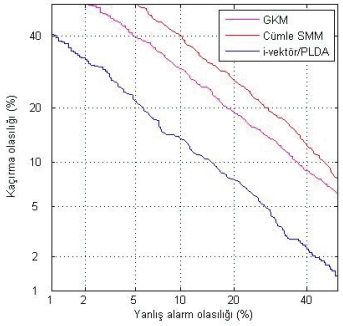
Şekil 5 ve Tablo 1'de görüldüğü gibi en iyi konuşmacı tanıma sonuçları cümle SMM yöntemi ile elde edilmiştir. Bu sonuç SMM yönteminin artikülasyon bilgisinden daha iyi faydalanmasına bağlanabilir. Ayrıca, GKM ve cümle SMM yöntemleri mobil veri tabanında son derece başarılı sonuçlar vermiştir. Bu durum veri tabanında sadece iki telefon tipinin bulunmasına ve bu telefonların mikrofon kalitelerinin benzer olmasına bağlanabilir. Önceki çalışmalarımızda daha fazla kanal durumu içeren veri tabanları ile yaptığımız testlerde bu yöntemler daha yüksek hata oranları vermiştir (Buyuk, 2011).

i-vektör/PLDA yönteminin hata oranı diğer iki yöntemden oldukça yüksektir. Bu durum veri tabanındaki kanal çeşitliliğinin azlığı ve eğitim/test cümlelerinin kısalığından kaynaklanıyor olabilir. Önceki çalışmalarda, metne bağımlı uygulamalar için i-vektör/PLDA tarzı yöntemlerin direkt uygulanması metinden bağımsız uygulamalardakine benzer performans artışını vermemiştir (Aronowitz, 2012; Stafylakis vd., 2013; Kenny vd., 2014). Bununla birlikte, metne bağımlı uygulamalarda i-vektör/PLDA yönteminin kısıtlı ses içeriğinden faydalanacak şekilde iyileştirilmesi önemli performans artışları sağlamıştır (Larcher vd., 2013; Novoselov vd., 2014). Bu konuyla ilgili çalışmalarımız devam etmektedir.

Tablo 2 ve Şekil 6'da kayıttan sahte doğrulama deneyi sonuçları verilmiştir. Kayıttan sahte doğrulama testleri sırasında normal sahte doğrulama denemeleri kayıttan sahte doğrulama saldırıları ile değiştirilmiştir. Bu nedenle bu deneydeki sahte doğrulama deneme sayısı önceki deneyden farklıdır. Tablo 2'deki sonuçlar elde edilirken 1462 gerçek ve 1462 sahte doğrulama denemesi yapılmıştır. Şekil 6'da kayıttan sahte doğrulama deneyi tanıma hata oranı eğrileri görülebilir.

Tablo 2: Kayıttan sahte doğrulama deneyi yüzde eşit hata oranları (%EHO).

	%EHO
GKM	19.357
CÜMLE SMM	23.734
İ-VEKTÖR/PLDA	11.764



Şekil 6: Kayıttan sahte doğrulama deneyi tanıma hata oranı eğrisi.

Tablo 2 ve Tablo 1'deki sonuçlar karşılaştırıldığında üç metodun da sahte doğrulama saldırılarından büyük ölçüde etkilendiği gözlenmektedir. Hata oranları yaklaşık %0.5-1 aralığından %10-25 aralığına çıkmaktadır. Kayıttan sahte doğrulama ataklarından en az etkilenen yöntem i-vektör/PLDA'dir. En fazla etkilenen yöntem ise cümle SMM'dir.

(Wu vd., 2014)'te İngilizce bir MBTC uygulamasında kayıttan sahte doğrulama ataklarının etkisi incelenmiştir. Bu çalışmada sadece GKM ve SMM tabanlı yöntemler test edilmiştir. Testlerde, GKM ve SMM yöntemleri normal sahte doğrulama denemeleriyle bayan konuşmacılar için sırasıyla %2.39 ve %3.67 eşit hata oranı vermiştir. Hata oranları kayıttan sahte doğrulama ataklarıyla %20.05 ve %21.95'e yükselmiştir. Bu çalışmada elde edilen sonuçlar bizim aynı yöntemlerle elde ettiğimiz sonuçlara oldukça yakındır. Kayıttan sahte doğrulama deneyinde hata oranının beklenen düzeyin bir miktar altında kalması daha önce bahsedilen deney düzeneğinden kaynaklanıyor olabilir. Gelecekte, deney düzeneğini orijinale daha yakın kayıtlara elde edecek şekilde değiştirerek sahte doğrulama atakları gerçekleştirilecektir. Ayrıca sahte doğrulama ataklarındaki yüksek geri plan gürültüsü de hata oranlarını etkilemektedir. Yeni deneylerde geri plan gürültüsü üzerinde de durulacaktır.

Bu çalışmada elde edilen sonuçlardan anlaşılacağı gibi mobil tek cümle uygulaması kayıttan sahte doğrulama ataklarına karşı son derece kırılgandır. Mobil platformlarda güvenilir konuşmacı tanıma sistemi gerçekleştirmek için kayıttan sahte doğrulama ataklarına karşı önlemler alınması gerekmektedir.

Sonuçlar ve Tartışma

Bu çalışmada mobil konuşmacı tanıma uygulamalarında kayıttan sahte doğrulama atakları incelenmiştir. Bu amaçla iki farklı akıllı telefon kullanılarak yeni bir metne bağımlı tek cümle veri tabanı oluşturulmuştur. Bu veri tabanı ile yaptığımız testler kayıttan sahte doğrulama saldırılarının konuşmacı tanıma performansını son derece olumsuz etkilediğini göstermiştir. Günümüzde herhangi bir kullanıcının parola cümlesini kaydetmenin kolaylığı düşünüldüğünde, bu saldırılara karşı etkili önlemlerin alınmasının gerekliliği daha iyi anlaşılacaktır.

Literatürde sahte doğrulama ataklarına karşı geliştirilen yöntemler, saldırıların doğrulama öncesinde tespit edilmesi üzerine

yoğunlaşmıştır. Kayıttan sahte doğrulama ataklarını belirleme konusunda, sisteme önceden bırakılmış kayıtlarla karşılaştırma temeline dayanan yöntemler bulunmaktadır. Büyük çağrı merkezlerine kurulan konuşmacı tanıma sistemlerine saatte binlerce doğrulama isteği gelmektedir. Bu yoğun kullanım düşünüldüğünde önceki kayıtlarla karşılaştırma temeline dayanan yöntemlerin pratikte uygulanma şansı oldukça azdır. Gelecekteki çalışmalarımız kayıttan sahte doğrulama ataklarının belirlenmesi için yeni yaklaşımların önerilmesi konusunda yoğunlaşacaktır.

Teşekkür

Bu çalışma TÜBİTAK 3001 programı çerçevesinde 114E742 numaralı proje kapsamında desteklenmiştir.

Kaynaklar

Alam, M.J., Kenny, P., Bhattacharya, G. Stafylakis, T., (2015). Development of CRIM System for the Automatic Speaker Verification Spoofing and Countermeasures Challenge 2015, Proc. of the European Conference on Speech Communication and Technology 2015 (INTERSPEECH 2015).

Alegre, F., Janicki, A., Evans, N., (2014). Re-assessing the threat of replay spoofing attacks against automatic speaker verification, in Proc. Int. Conf. of the Biometrics Special Interest Group (BIOSIG), 2014.

Aronowitz, H., (2012). Voice biometrics for user authentication, Afeka-AVIOS Speech Processing Conference 2012, Tel-Aviv, Israel, pp. 1-4.

Blouet, R., Mokbel, C., Mokbel, H., Soto, E. S., Chollet, G., Greige, H., (2004). Becars: A free software for speaker verification, Proc. of the Speaker and Language Recognition Workshop 2004 (ODYSSEY 2004), Toledo, Spain.

Buyuk, O., (2011). Telephone-based Text-Dependent Speaker Verification, PhD. Thesis, Bogazici University, Turkey.

Buyuk, O., Arslan, L.M., (2012). Model Selection and Score Normalization for Text-Dependent Single Utterance Speaker Verification, Turkish Journal of Electrical Engineering and Computer Sciences 20 (sup.2), 1277-1295.

Chen, N., Qian, Y., Dinkel, H., Chen, B., Yu, K., (2015). Robust Deep Feature for Spoofing Detection - The SJTU System for ASVspoof 2015 Challenge, Proc. of the European Conference on Speech Communication and Technology 2015 (INTERSPEECH 2015).

Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., Ouellet, P., (2011). Front-end factor analysis for speaker verification, IEEE Transactions on Audio, Speech, and Language Processing 19 (4), pp. 788-798.

Ferrer, L., McLaren, M., Scheffer, N., Lei, Y., Graciarena, M., Mitra, V., (2013). A noise-robust system for NIST 2012 speaker recognition evaluation, Proc. of the European Conference on Speech Communication and Technology 2013 (INTERSPEECH 2013), Lyon, France, pp. 1981-1985.

Garcia-Romero, D., Espy-Wilson, C. Y., (2011). Analysis of i-vector length normalization in speaker recognition systems, Proc. of the European Conference on Speech Communication and Technology 2011 (INTERSPEECH 2011), Florence, Italy, pp. 249-252.

Hasan, T., Sadjadi, S. O., Liu, G., Shokouhi, N., Boril, H., Hansen, J. H., (2013). CRSS systems for 2012 NIST speaker recognition evaluation, Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing 2013 (ICASSP 2013), Vancouver, Canada, pp. 6783-6787.

Janicki, A., (2015). Spoofing Countermeasure Based on Analysis of Linear Prediction Error, Proc. of the European Conference on Speech Communication and Technology 2015 (INTERSPEECH 2015).

Kenny, P., (2010). Bayesian speaker verification with heavy-tailed priors, Proc. of the Speaker and Language Recognition Workshop 2010 (ODYSSEY 2010), Brno, Czech Republic, pp. 014.

Kenny, P., Stafylakis, T., Alam, J., Oullet, P., Kockmann, M. (2014). Joint factor analysis for text-dependent speaker verification, Proc. of the Speaker and Language Recognition Workshop 2014 (ODYSSEY 2014), Joensuu, Finland, pp. 200-207.

Larcher, A., Lee, K. A., Ma, B., Li, H. (2013). "Phonetically constrained PLDA modeling for text-dependent speaker verification with multiple short utterances", Proc. of the IEEE International

Conference on Acoustics, Speech and Signal Processing 2013 (ICASSP 2013), Vancouver, Canada, pp. 7673-7677.

Novoselov, S., Pekhovsky, T., Shulipa, A., Sholokhov, A., (2014). Text-dependent GMM-JFA system for password based speaker verification, Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing 2014 (ICASSP 2014), Florence, Italy, pp. 729-737.

Prince, S. J. D., Elder, J. H., (2007). Probabilistic linear discriminant analysis for inferences about identity, Proc. of the IEEE International Conference on Computer Vision 2007 (ICCV 2007), Rio de Janeiro, Brazil, pp. 1-8.

Reynolds, D.A., Quatieri, T.F., Dunn, R.B., (2000). Speaker Verification Using Adapted Gaussian Mixture Models, Digital Signal Processing, 10 (1-3), 19-41.

Sadjadi, S. O., Slaney, M., Heck, L. P., (2013). MSR identity toolbox: A MATLAB toolbox for speaker recognition research, version 1.0, Technical Report, Microsoft Research, Conversational Systems Research Center (CSRC), Nov. 2013.

Shang, W., Stevenson, M., (2010). Score normalization in playback attack detection, Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing 2010 (ICASSP 2010).

Stafylakis. T., Kenny, P., Ouellet, P., Perez, J., Kockmann, M., Dumouchel, P., (2013): I-Vector/PLDA variants for text-dependent speaker recognition, Technical Report, June 2013, Montreal, CRIM.

Sturim, D., Campbell, W., Dehak, N., Karam, Z., McCree, A., Reynolds, D. A., Richardson, F., Torres-Carrasquillo, P., Shum, S., (2011). The MIT LL 2010 speaker recognition evaluation system: Scalable language-independent speaker recognition, Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing 2011 (ICASSP 2011), Prague, Czech Republic, pp. 5272-5275.

Super Monitoring (2013), State of Mobile 2013, http://www.supermonitoring.com/blog/2013/09/23/state-of-mobile-2013-infographic/#tt Son erişim tarihi: 19 Mart 2014

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P., (2006). The HTK Book (for HTK Version 3.4), Cambridge University Engineering Department.

Wu, Z., Kinnunen, T., Chng, E.S., Li, H., Ambikairajah, E., (2012). A study on spoofing attack in state-of-the-art speaker verification: the telephone speech case, Proc. of the Asia-Pacific Signal Information Processing Association Annual Summit and Conference 2012 (APSIPA ASC 2012).

Wu, Z., Gao, S., Cling, E. S., Li, H, (2014). A study on replay attack and anti-spoofing for text-dependent speaker verification, Proc. of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference 2014 (APSIPA ASC 2014).

Wu, Z., Evans, N., Kinnunen, T., Yamagishi, J., Alegre, F., Li, H (2015a). Spoofing and countermeasures for speaker verification: a survey, Speech Communication 66, pp. 130–153.

Wu, Z., Kinnunen, T., Evans, N., Yamagishi, J., Hanilci, C., Sahidullah, M., Sizov, A., (2015b). ASVspoof 2015: the First Automatic Speaker Verification Spoofing and Countermeasures Challenge, Proc. of the European Conference on Speech Communication and Technology 2015 (INTERSPEECH 2015).

The vulnerability of mobile text-dependent single utterance speaker verification to replay attacks

Extended abstract

Adapting different technologies for mobile platforms have become an important industry due to the vast use of mobile applications. With the significant increase in mobile applications, the security issues have also become a major concern for the mobile users. The aim of speaker recognition is to recognize the identity of the speaker from his/her voice. Thus, it provides a good alternative for mobile security. Speaker recognition technology can be used to increase the overall security of the applications requiring high security. It can also add extra security to an application by verifying the user with the voice in addition to a typed password.

Speaker verification applications might be divided into two categories; text-dependent and text-independent. In text-dependent applications, vocabulary is usually constrained to digit strings or pre-defined pass phrases. In text-independent applications, there is no such constraint and system tries to verify the identity of the speaker from his/her natural speech. In text-dependent single utterance (TDSU) speaker verification, speakers repeat a fixed pass phrase in both enrollment and authentication sessions. The repetition of a single utterance improves the overall recognition accuracy of the system since the authentication utterance is included in the enrollment as a whole. Repetition of the same utterance also makes the usage easier. However, TDSU applications become vulnerable to replay attacks due to the same reason. A pre-record of the pass phrase might be used to spoof the system. In this study, we evaluate the robustness of mobile TDSU applications to replay attacks.

In order to test the robustness of mobile TDSU applications to replay attacks, we construct a new speaker recognition database. We choose the Turkish utterance "benim parolam ses kaydımdır (my voice is my password)" as the pass phrase in the TDSU task since it contains 5 of the 8 vowels in the Turkish language. The database consists of 124 speakers. 62 of the speakers are female and 62 are male. The recordings are taken in 2 separate sessions using 2 different smart phones. Using the database, a realistic simulation of the replay attacks is performed by playing the recordings from one phone and recording to the other. The replay recordings are used as imposter trials in the verification tests.

Until recently, Gaussian mixture models (GMMs) have been the dominant modeling approach for text independent speaker verification. In GMM, each speaker is modeled with a mixture of Gaussians. Generally, speaker models are adapted from a speaker independent universal background model (UBM). Maximum a posterior (MAP) method is usually used for the adaptation. In text dependent applications, hidden Markov model (HMM) based approaches are used since they better capture the co-articulation information. In a TDSU task, a single whole phrase HMM might be constructed for the pass phrase. The sentence HMM topology might be preferred over the phone HMM in order to better model the co-articulation and improve the verification performance. Recently, very powerful channel compensation techniques such as joint factor analysis (JFA), i-vector and i-vector/probabilistic linear discriminant analysis (i-vector/PLDA) are proposed. The methods achieved very good verification performance especially for text independent tasks. The performance gain of the methods for the text-dependent tasks is still investigated.

In this study, we implement GMM, sentence HMM and i-vector/PLDA methods for the TDSU speaker verification task. The methods are tested against the replay spoofing attacks. The baseline equal error rate (EER) of the three methods with zero-effort imposter trials are about 0.5-1%. The best performance is achieved with the sentence HMM method in the baseline case. The verification performance of all three methods significantly decreases when zero-effort imposter trials are replaced with the replay spoofing attacks. The equal error rate increase to 10-25% from 0.5-1% with the replay trials. i-vector/PLDA results in the best performance in the spoofing experiment.

Keywords: *Speaker verification, text dependent single utterance, replay attacks, mobile devices, smart phones.*