Robust Voice Liveness Detection and Speaker Verification Using Throat Microphones

Md Sahidullah, *Member, IEEE*, Dennis Alexander Lehmann Thomsen, Rosa Gonzalez Hautamäki, Tomi Kinnunen, *Member, IEEE*, Zheng-Hua Tan, *Senior Member, IEEE*, Robert Parts, Martti Pitkänen

***Abstract*—While having a wide range of applications, automatic speaker verification (ASV) systems are vulnerable to spoofing attacks, in particular, replay attacks that are effective and easy to implement. Most prior work on detecting replay attacks uses audio from a single acoustic microphone (AM) only, leading to difficulties in detecting high-end replay attacks close to indistinguishable from live human speech. In this paper, we study the use of a special body-conducted sensor, throat microphone (TM), for combined voice liveness detection (VLD) and ASV in order to improve both robustness and security of ASV against replay attacks. We first investigate the possibility and methods of attacking a TM-based ASV system, followed by a pilot data collection. Secondly, we study the use of spectral features for VLD using both single-channel and dual-channel ASV systems. We carry out speaker verification experiments using Gaussian mixture model with universal background model (GMM-UBM) and i-vector based systems on a dataset of 38 speakers collected by us. We have achieved considerable improvement in recognition accuracy, with the use of dual-microphone setup. In experiments with noisy test speech, the false acceptance rate (FAR) of the dual-microphone GMM-UBM based system for recorded speech reduces from 69.69% to 18.75%. The FAR of replay condition further drops to 0% when this dual-channel ASV system is integrated with the new dual-channel voice liveness detector.**

***Index Terms*—Automatic speaker verification, anti-spoofing, voice liveness detection, two-channel countermeasure, replay attack, throat microphone**

I. Introduction

Speech, as one of the most information-rich biosignals, is the primary means of human communication. Besides the message relayed through spoken words, the speech signal conveys information of the speaker's identity, enabling recognition of the person both by human listeners and *automatic speaker verification* (ASV) techniques [1], [2]. While the traditional role of ASV has been in assisting forensic speaker comparison and surveillance, there is an increasing interest to use ASV for user authentication in consumer applications including smartphone log-in [3], e-commerce [4], mobile banking [5] and physical access control[1].

In contrast to other biometric technologies, such as fingerprint and face recognition, ASV has a considerably wider range of potential applications since it requires no additional hardware investments: a speech signal can be acquired through the native communication channel of a particular application via landline phone, conventional cellular phone, radio phone, satellite phone, smartphone, tablet, or a desktop PC with a headset, to name a few. Applications range from tactical use in military and police enforcement to telephone banking, stockbroking, telephone calls and teleconferences. Many of these require highest possible level of security and privacy that can be enhanced using ASV technology. For instance, in a telephone banking scenario, the operator can use an ASV system to verify the caller's identity before disclosing sensitive information, such as bank account balance.

Unfortunately, similar to other biometric identifiers [6], ASV systems are vulnerable to so-called *presentation attacks* or *spoofing attacks* involving malicious effort to misguide the ASV system so that the attacker would be falsely accepted as another targeted speaker. The (currently known) spoofing attacks against ASV systems fall into one of four main categories, *impersonation* [7], [8], *replay* [9], [10], *speech synthesis* and *voice conversion*. A detailed survey of the vulnerabilities is provided in [11] and references therein. In this study, we focus exclusively on replay attacks — playback of a pre-recorded target speaker's voice sample to the ASV system sensor —- that have remained comparatively much less studied [12], [13], [14], [15], despite the apparent ease to implement them; while constructing state-of-the-art speech synthesis and voice conversion attacks requires considerable expert knowledge, replay attacks could, in principle, be executed by *anyone* using a consumer device with a loudspeaker and a device to play audio files — such as a smart-phone.

A. Replay attack countermeasures, their pros and their cons

To increase general trust to the security of ASV systems, continued quest for *countermeasures* to detect and reject spoofing attacks, especially replay, is critically important. There are a number of ways to address this problem. The first one, based on a challenge-response approach, is *utterance verification* [16]: a failure to produce a text prompted by

Md Sahidullah, Rosa Gonzalez Hautamäki and Tomi Kinnunen are with the University of Eastern Finland, Joensuu, FI-80101, Finland. e-mail: sahid@cs.uef.fi, rgonza@cs.uef.fi, tkinnu@cs.uef.fi.

Dennis Thomsen and Zheng-Hua Tan are with Aalborg University, 9220 Aalborg, Denmark. e-mail: zt@es.aau.dk, dalth@es.aau.dk.

Robert Parts and Martti Pitkänen are with APLcomp, Helsinki 00170, Finland. e-mail: parts@neti.ee, martti.pitkanen@aplcomp.fi.

The paper reflects some results from the OCTAVE Project (#647850), funded by the Research European Agency (REA) of the European Commission, in its framework programme Horizon 2020. The views expressed in this paper are those of the authors and do not engage any official position of the European Commission.

Manuscript submitted Dec 15, 2016.

[1]The recently concluded OCTAVE project https://www.octave-project.eu/ addressed both logical and physical access control scenarios.

the system is an indication of a potential replay attack. Secondly, as human can never produce the exact same speech signal, some countermeasures use *template matching* or *audio fingerprinting* to verify whether the input utterance was presented to the system earlier [17], [11]. Thirdly, similar to the detection of speech synthesis and voice conversion attacks in the ASVspoof challenge [18], some work looks into *statistical acoustic characterization* of authentic and replay speech. As replayed speech differs from an authentic speech through the use of digital-to-analog converters, amplifiers, loudspeakers and reverberation, these operations leave out acoustic traces, enabling detection of replay attacks using spectral cues [19], [20]. Finally, a fourth class of methods uses *voice liveness detection* (VLD); for instance, in [21], *pop noise* present in live human speech recorded with a microphone without a pop shield but absent from replay recordings, was used to develop a countermeasure based on two microphones.

In the same way as no single biometric modality is superior to others, due to the complicated trade-offs [22] between accuracy, user experience, speed, cost and other factors, no perfect ASV spoofing countermeasure exists. All the four types of countermeasures listed above offer some protection but have certain limitations, too. Firstly, utterance verification or text-prompting approach could be circumvented by recognition of the prompted text followed by the use of voice conversion or speech synthesis techniques.

Secondly, the audio fingerprinting approach is based on the premise that the test speech sample is an imperfect replica of the enrollment (or earlier) test utterances. Thus, it may fail to detect replay samples that are strategically modified for the purpose of avoiding to be detected, but still being accepted as the target speaker. Most obviously, this could be done simply by acquiring *another* rendition of the target speaker pass-phrase, perhaps prepared using unit selection techniques from publicly available speech of the target speaker (lectures, public speeches or video uploads to social media). Alternatively, one might introduce purposeful local time-frequency modifications, such as time jitter or frame dropping during non-speech regions. Another issue with audio fingerprinting would be the dynamically increasing database size with the stored "known" utterances (or their audio fingerprints), causing potential scalability issues.

Thirdly, the statistical characterization approach may fail if the replay sample is collected with high-quality devices in a clean environment[2], or when the replay artifacts might be confusable with other, naturally occurring nuisance factors due to speech coding (telephony) or natural room reverberation, for instance. Indeed, the results from the recent ASVspoof 2017 challenge [23] outline the notoriously challenging nature of such countermeasures, especially regarding generalization across different replay environments. Finally, the liveness detection approach, while potentially very accurate, requires an additional sensor (here, a throat microphone) and therefore may not be applicable in all application scenarios, but could be very useful where a very high level of security is required.

[2]The limiting case would be an exact digital copy of a target speaker's recording — it cannot be detected by *any* such detector.

In the absence of any published comprehensive work that compares all the four types of countermeasures, and without published technical details of the spoofing countermeasures used in commercial ASV systems, the authors take an agnostic view on superiority claims of any single countermeasure over another one. There are, however, two driving motivations why we focus on voice liveness detection using additional sensors. Firstly, from the four types of ASV countermeasures, it has received least attention. Secondly, encouraged by the impressive performance improvements due to throat microphone in general ASV tasks [24], [25], [26], including the authors' preliminary recent work [27], we got inspired to look further into the question whether throat microphones might be useful source for spoofing countermeasures as well. The focus of this work, therefore, is on developing a novel VLD system that uses joint characteristics of the signals collected using acoustic and throat microphones to distinguish live human voice from replay samples, and specifically, under adverse conditions. Similar to the recently concluded ASVspoof 2017 challenge [23], we restrict our focus on text-dependent ASV, the most relevant case in authentication applications.

B. Our contribution: both robust and secure ASV with a replay attack countermeasure using throat microphones

Most replay countermeasures utilize audio data from a single sensor only, *i.e.* an acoustic microphone (AM). By drawing inspiration from the successful use of additional sensors in other biometric tasks [28], [29], [30], [31] besides the primary sensor, we propose to use a specific skin-attached non-acoustic sensor [32], *throat microphone* (TM), or *laryngophone*, to enhance voice liveness detection. The work closest in spirit to ours is [21] that also uses dual-channel voice liveness detection based on the pop noise. Different from [21] that uses two homogenous acoustic microphones, however, our other microphone is the contact microphone as shown in Fig. 1.

The overall goal, largely stemming from a recent industry-driven OCTAVE project, is to enhance the *robustness* of both ASV performance *and* voice liveness detection under adverse conditions. Indeed, besides the recently highlighted problem of strong performance dependency on the corpus [33], similar to results seen in the domain of face recognition [34], recent independent studies in [35] and [36] have revealed severe sensitivity of the state-of-the-art statistical spoofing detectors to even modest amounts of additive noise[3]. In short, most findings reported on high-quality spoofing corpora or across different corpora do not appear to translate well to adversarial ASV scenarios or domain (data) mismatch.

Now, since the throat microphone is by-design more robust against background noise compared to the acoustic microphone (reviewed in Section II), it provides a viable additional signal source for more robust and secure ASV operation under both clean and noisy conditions. For this reason, throat microphones have been used specifically in military, aviation,

[3]Even if [35] and [36] addressed noise sensitivity of a different spoofing task, detection of voice conversion and synthetic speech (taken from the ASVspoof challenge [18]), similar countermeasures are typically adopted with minor modifications to detect other types of attacks, and therefore we expect similar degradations in the case of replay attacks.

Fig. 1: Two of the throat microphones used in this study.

law enforcement, sports and other similar scenarios where the subjects wear helmets, masks or full-face breathing apparatuses. Although the use of TMs has been widely explored in various speech processing tasks, the use in ASV [24], [25], [26], [37], [38] is scarce.

To the best of our knowledge, this is the first work that uses throat microphone for voice anti-spoofing, in particular for voice liveness detection. It is of fundamental scientific interest to find out whether throat microphones may provide additional protection not provided by the existing countermeasures relying on acoustic microphones. Enabled by the publicly available spoofing evaluation data through the ASVspoof [39], [23] and the AVspoof [19] challenges, much of the recent work has been devoted on developing advanced spoofing countermeasures to protect ASV systems from artificial speech and replay attacks. But since these activities have solely been focused on the single-channel acoustic microphone countermeasures, an accurate picture of the potential of auxiliary sensors for voice anti-spoofing — here, a throat microphone — is thus far missing. The purpose of this study is to fill some of that void.

Since neither publicly available evaluation data nor commonly agreed, multiple-time validated countermeasures for dual-channel ASV countermeasures exist, we define and collect our custom data. This work extends our preliminary work [27] in several respects. Firstly and foremost, [27] addressed only baseline ASV (zero-effort impostors) without any spoofing attack considerations, which forms the primary focus of this work. Secondly, as there is no existing work for the study of replay attacks against TM, we define such a replay attack evaluation scenario, including novel attacks that involve a physical contact of the throat microphone to various different attack devices. Specifically, besides conventional loudspeakers, we include replay attacks against TM using an *audio exciter* device. Differently from most reported work on replay attack assessments in the ASV context (prior to the ASVspoof 2017 challenge [23]), we collect simulated replay attacks at multiple recording sites to account for the potential domain dependency of results on the environment. Thirdly, besides just the usage of TM for speech activity detector to aid ASV under noisy conditions [27], we adopt a potentially more robust technique by combining spectral information across the AM and the TM sensors using *ideal ratio masks* (IRMs) [40]. Even if IRMs were used in other robust speech processing tasks [41], [42], [43], we are unaware of their prior use in the context of dual-channel ASV tasks. Finally, we provide a comparative assessment of two commonly used fusion strategies, score fusion and feature fusion, for the task of voice liveness detection. The purpose of that analysis is to gain insight as to whether the two streams — AM and TM — should be treated as independent, or whether the synchronized cross-microphone information (with increased feature dimensionality) provides a more viable starting point for replay attack countermeasures.

II. Throat microphone: an alternative non-acoustic microphone

A. Background and motivation

The conventional acoustic microphone picks not only the target signal but the background noise. This has inspired the development of many non-acoustic vibration-based sensors to acquire speech signals more robustly [44]. For example, *physiological microphone* (PMIC) with piezo-electric crystal sensor is often used [32] to capture neck-skin vibration due to speech production and to subsequently convert this into an electric signal. Other than PMIC, *bone microphone* and *in-ear microphone* are also used [45]. Besides those vibration sensors, electromagnetic sensors such as *electroglottograph* (EGG) and *general electromagnetic motion systems* (GEMS) are also used. For an extensive review of such sensors, we point the interested reader to [45].

From the various alternative non-acoustic sensors, we focus solely on the throat microphone. The primary reasons for this choice are technological maturity, cost efficiency and noise robustness. As for the maturity, the history of throat microphones dates back to the advent of the second world war (WW2), the US patent of [46] being possibly the first reported work. In WW2, throat microphones were used especially by the German pilots to improve communication intelligibility in the noisy warplane cockpits. Since then, they have been deployed commercially especially in radio phone communication to maintain intelligible conversation even under extreme conditions. Nowadays, many high-quality throat microphones are available at low cost off-the-shelf from various manufacturers for modern end-user devices [47], [48], [49], including smart-phones. This enables cost-effective integration of throat microphones into modern ASV use cases enlisted in Section I. Moreover, as the sensor trunk can be positioned quite comfortably and steadily, the throat mic can be furnished with an efficient power source, which is one of the main stumbling blocks of wearable technology. While the use of an additional microphone would obviously decrease user convenience in certain applications, they are particularly deployable in applications where the subject wears helmets or respiratory protection devices natively. Our primary applications would be in office or home environments for such teleconference or specific e-banking where a very high level of trust to the other conversion party is required.

B. General Use of Throat Mics in Speech Processing

Techniques for robust speech recognition and speech activity detection (SAD) in highly noisy, non-stationary environ-

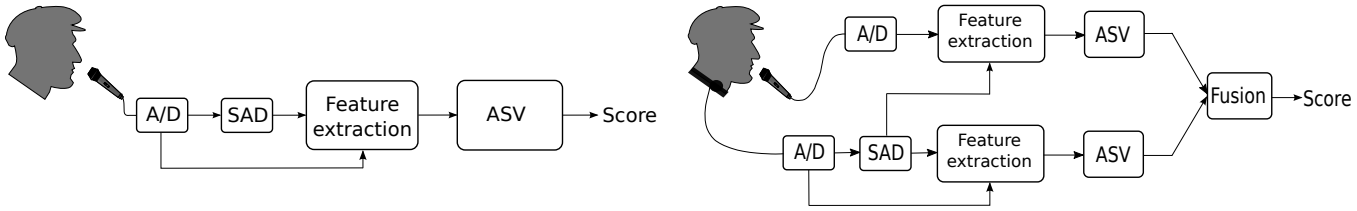


Fig. 2: Conventional speaker verification setup with a single acoustic microphone (*left*) and the combined use of acoustic and throat microphones (*right*).

ments using several heterogeneous sensors have been studied in [50]. The hardware prototypes integrate AM with bone microphones and TM among others into headsets. Another wearable recording system used in [51] integrates a close-talking, a monophonic far-field, and a TM in addition to a 4-channel far-field microphone array to create a multi-channel database for speech recognition. In [52], a technique for estimating clean acoustic speech features by combining TM and AM recordings using a probabilistic optimum filter (POF) mapping is proposed for speech recognition. Since TM speech is relatively less noisy compared to AM, when captured in adverse environments, it can be used to detect speech regions accurately. In [53] and [54], this idea is used and improved recognition performance is obtained when TM speech is used for SAD. In [55], various adaptation methods such as maximum likelihood linear regression and sigmoid low-pass filtering are studied in the context of whispered speech recognition with the help of TM signal. In [56] and [57], TM is used for voice quality assessment. Throat microphone signals are also used in speech enhancement as clean reference signal for the objective performance measure for speech enhancement algorithms [58]. Though TM speech is less affected by the ambient noise, its intelligibility is lower than AM speech [59]. For this reason, sometimes the quality of TM speech also needs to be improved. In [60], a phone-dependent Gaussian mixture model-based statistical mapping have been explored for this purpose to construct probabilistic mappings between TM and AM speech signals. Various spectral mapping techniques are compared in [61] for the enhancement of TM speech.

C. Use in Automatic Speaker Verification

In spite of its high robustness against environmental noise, throat microphone are studied scarcely in the ASV context. The previous studies in throat microphone based speaker recognition used auto-associative neural network (AANN) for modeling target speakers [24], [25], [26]. Performance was evaluated for closed-set speaker identification task. In our recent work [27], we studied the performance of acoustic and throat microphone based ASV using more modern GMM-UBM [62] and i-vector [63] based speaker recognition, including the use of TM as a side-information to derive robust SAD labels and score fusion of the AM and TM signals to combine speaker cues across the two channels. The overall system diagram for conventional acoustic microphone and proposed throat-microphone based dual microphone is shown in Fig. 2.

In [27], we used the TM to obtain active speech regions for the AM signal. In this study, we consider a further alternative technique to benefit from the dual-channel information. We apply so-called *ideal ratio mask* (IRM) [40] to weight each frequency bin of AM, before SAD. Let the short-term Fourier power spectrum of t-th speech frame for AM and TM be $|X_{\mathrm{AM}}(t,f)|^2$ and $|X_{\mathrm{TM}}(t,f)|^2$. Here, f is the discrete frequency bin. To this end, we apply the following weighting:

$$\mathrm{IRM}(t,f) = \left(\frac{|X_{\mathrm{TM}}(t,f)|^2}{|X_{\mathrm{AM}}(t,f)|^2}\right)^{1/\beta}, \qquad (1)$$

where $\beta \neq 0$ is a control parameter. Then, we extract features (such as MFCCs) from the weighted power spectrum $\mathrm{IRM}(t,f) \times |X_{\mathrm{AM}}(t,f)|^2$. The special case $\beta = 1$ means extracting features from TM only (the AM spectrum cancels out), while $\lim_{\beta\to\infty} \mathrm{IRM}(t,f) = 1$ ensures that large values of β imply feature extraction using the acoustic signal only. All the other values, $1 < \beta < \infty$ correspond to "blending" the information across the two mics.

III. Voice Liveness Detection Using Throat Microphones

If an attacker deploys a high-end loudspeaker for replay, the signal captured by an acoustic microphone might be close to indistinguishable from a live human voice, thereby making it *impossible* to detect the replay attack relying solely on acoustic cues. This is where the throat microphone comes into help: since a live human wears the throat microphone in his/her neck, the frequency characteristics of this signal differs from a replayed represented to the throat microphone. This is due to differing conduction properties of the human tissue versus the acoustic transfer properties of the loudspeaker and the room. This is illustrated in Fig. 3 which displays the long-term average spectra (LTAS) of AM and TM speech of three different sentences for both live human and replay conditions. The replay recordings of the acoustic microphone are similar to the playback recording, especially at the low and middle frequency range, while the spectra for the TM signal are evidently different in original and playback recording.

Further, Fig. 5 displays spectrograms of a live human speech and two replay signals from two different configurations (sessions). The replayed signal of the acoustic microphone signal (Fig. 5(b) and Fig. 5(c)) are similar to the live human spectrogram (Fig. 5(a)). But for the throat microphone, the replay recordings (Fig. 5(e) and Fig. 5(f)) are clearly distinguishable from the original recording (Fig. 5(d)). Hence, we

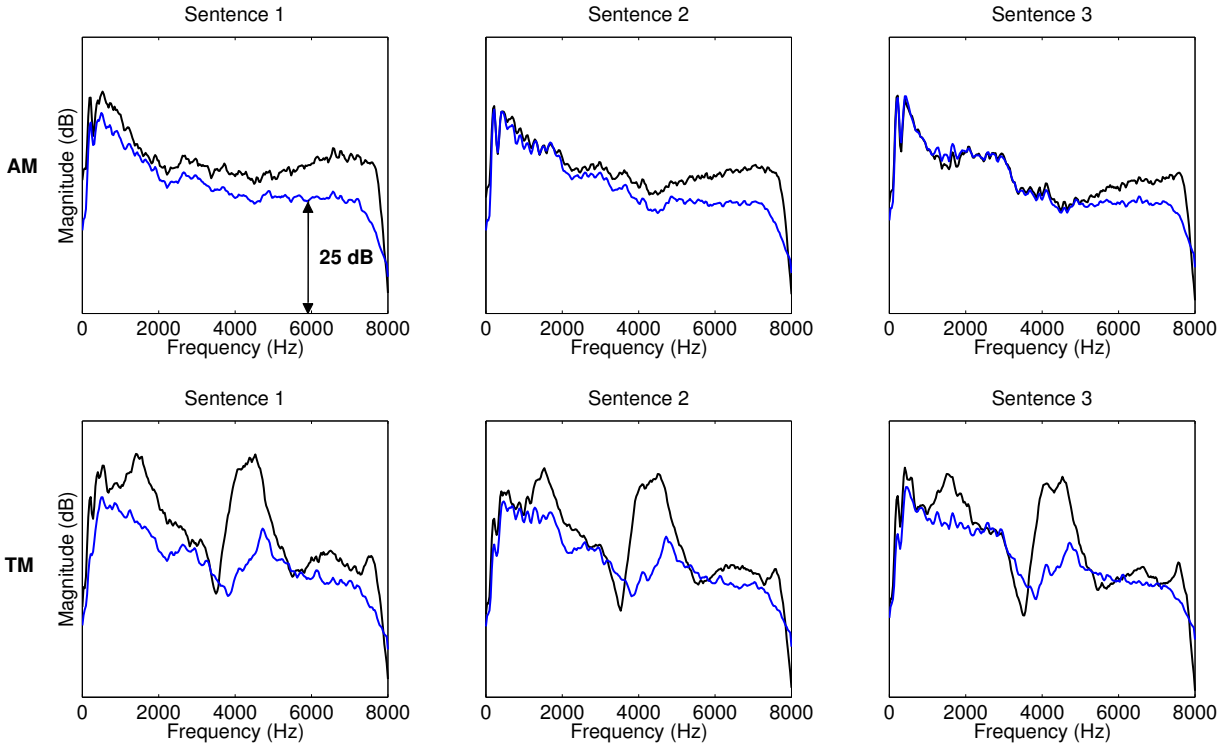


Fig. 3: Illustration of LTAS for AM and TM signal for same speaker and three sentences. The *black* line shows the LTAS of genuine recording. The *black* line shows the average of LTAS of replay recordings.

hypothesize that the spectral information acquired from the throat microphone signal could be more useful in discriminating live and replayed signal in comparison to the acoustic microphone.

A. Replay Attack Detection: State-of-the-Art Baseline

Speech-based spoofing countermeasures mostly employ spectral features as a front-end for speech representation [64] followed by a statistical classifier [65]. The standard *mel-frequency cepstral coefficient* (MFCC) features were used in many earlier spoofing detection studies [11]. A recent study on comparison of features [64] suggested, however, that the dynamic (delta and double delta) coefficients might be more useful for discriminating spoofed and real speech. In this study on ASVspoof 2015 corpus consisting voice conversion and speech synthesis based spoofed data, *linear frequency cepstral coefficients* (LFCCs) were found to provide the lowest overall detection error rates. In another recent study on the same corpus [20], the so-called *constant Q cepstral coefficient* (CQCC) features further outperformed LFCCs. In the current work under the context of dual-mic based replay attack, we adopt these three state-of-the-art features for the detection of replay attacks. We extract the features separately from both microphone channels.

Given an audio recording s, presented through any of the acoustic feature sets mentioned above, replay spoofing speech detection task is to decide whether s belongs to a genuine speech class (live human) — hypothesis $\mathcal{H}_0$, or a replay speech class — hypothesis $\mathcal{H}_1$. The decision is based upon a log-likelihood ratio score, $\Lambda = \log \mathrm{p}(\mathrm{s}|\mathcal{H}_0) - \log \mathrm{p}(\mathrm{s}|\mathcal{H}_1)$, where each of the two likelihoods are evaluated using Gaussian mixture models (GMMs) [66] trained using maximum likelihood. This approach was found to work consistently well on the ASVspoof 2015 challenge data [65].

B. Proposed dual-channel countermeasures

The above description of widely-used spoofing attack countermeasures for single-channel signals provides a strong methodological back-bone for generalizing the methods for dual-channels. To this end, we propose to use the joint feature based on features extracted separately from both the channels. The idea is to simply concatenate the features from both channels in frame-level. This will combine information from both the channels as well as will capture the inter-channel correlation of the speech features. Let $\mathbf{x}_{\mathrm{AM}}(t)$ and $\mathbf{x}_{\mathrm{TM}}(t)$ to denote the d-dimensional features at speech frame t, separately extracted from AM and TM channels, respectively. Then the $2d$-dimensional dual channel feature is given by,

$$\mathbf{x}_{\mathrm{DM}}(t) = \left[\mathbf{x}_{\mathrm{AM}}(t)^\top \quad \mathbf{x}_{\mathrm{TM}}(t)^\top\right]^\top. \tag{2}$$

For any data fusion method, an important question is in what ways the combined features or classifiers are complementary. Since our AM and TM signals are synchronous and since all

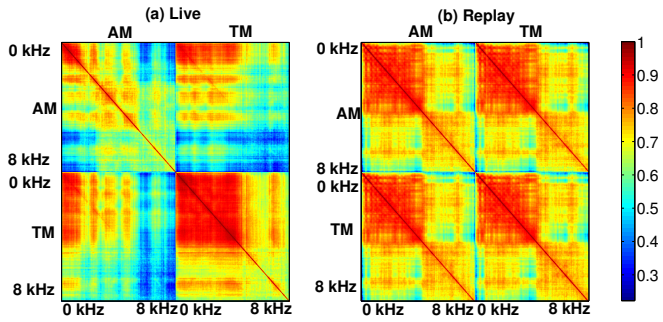


Fig. 4: Cross-channel correlation of AM and TM power spectrum for (a) live and (b) replay voice.

our feature extraction methods rely on power spectrum, we gain some insight by looking into spectral correlations across the two channels. To this end, consider a speech utterance with T frames and let $S_{\mathrm{AM}}(t, f_i)$ and $S_{\mathrm{TM}}(t, f_i)$ to denote the log-power spectral component of t-th frame at f_i frequency for AM and TM, respectively. The cross-channel power spectrum correlation can be expressed as,

$$\rho(f_i, f_j) = \frac{\frac{1}{T}\sum_{t=1}^{T} \tilde{S}_{\mathrm{AM}}(t, f_i)\tilde{S}_{\mathrm{TM}}(t, f_j)}{\sqrt{\frac{1}{T}\sum_{t=1}^{T}[\tilde{S}_{\mathrm{AM}}(t, f_i)]^2}\sqrt{\frac{1}{T}\sum_{t=1}^{T}[\tilde{S}_{\mathrm{TM}}(t, f_j)]^2}},$$

where $\tilde{S}_{\mathrm{AM}}(t, f_i) = S_{\mathrm{AM}}(t, f_i) - \mu_{\mathrm{AM}}(f_i)$ and $\tilde{S}_{\mathrm{TM}}(t, f_j) = S_{\mathrm{TM}}(t, f_i) - \mu_{\mathrm{TM}}(f_j)$ are the mean-subtracted log-power spectra of each channel. Here, $\mu_{\mathrm{TM}}(f_i)$ and $\mu_{\mathrm{TM}}(f_i)$ are mean of log-power spectral component at i-th frequency.

In Fig. 4, we display the cross-channel correlation matrix of AM and TM power spectrum for a randomly selected speech utterance (note that this also contains the within-channel correlations). We find that the cross-channel information is complementary for discriminating live human and replayed samples. Specifically, for the replay signal, the correlation between AM and TM signals is higher than the correlations of the two microphones for a live voice. This is because we collect replay signals by doing playback of the same acoustic signal. Note that this correlation information will be ignored if we treat the data from two channels separately, *e.g.*, score-level fusion of two systems.

IV. Experimental Setup

A. Collection of Speech Corpus

We collected the dual microphone speech data for text-dependent ASV setting in three different sites. The data was recorded synchronously in two channels with Scarlett 2i2 USB 2.0 audio interface manufactured by Focusrite[4] and all the recordings use a similar model of AM and TM. We recorded the samples using a web-based user interface with the Microsoft Edge web browser. The sampling frequency was set at 44.1 kHz. The phrases contained in the recordings are the same as the common phrases of Part I subcondition in the on-going RedDots initiative [67]. We record five different sessions for each subject, one noisy and four clean ones, in common office environments. In our ASV experiments, *matched condition* refers to the test case with clean speech whereas test with noisy session corresponds to the *mismatched condition*. We have collected speech signals in real-world noisy conditions for the mismatch scenario. Specifically, we collected all the data in casual home and office environments where one might use ASV technology when attending a secure teleconference or to access for instance tele-banking. Thus, the types of noises in our data contains background sounds from casual office or home environments including sounds from TV, coffee machine, babble noise, or sounds of furniture being moved around. In both matched and mismatched conditions, we use clean speech for training.

Our data consists of total 38 speakers, from which 30 (23 male and 7 female) are used for the ASV experiments while the remaining 8 are used for domain adaptation. The idea here is to use this limited throat microphone data to represent the acoustic space by means of domain adaptation. Three different clean sessions are used for training text-dependent speaker models. 10 different speaker models, corresponding to different phrases, are trained for each target speaker, yielding a total of 300 target models for our 30 speakers. The remaining two sessions, one clean and one noisy, are used for testing. Trials are designed so that the texts or *spoken-content* of a target model and test segment are identical. For each condition, there are 9000 trials, with 300 *genuine* or *target* trials and 8700 *impostor* or *nontarget* trials.

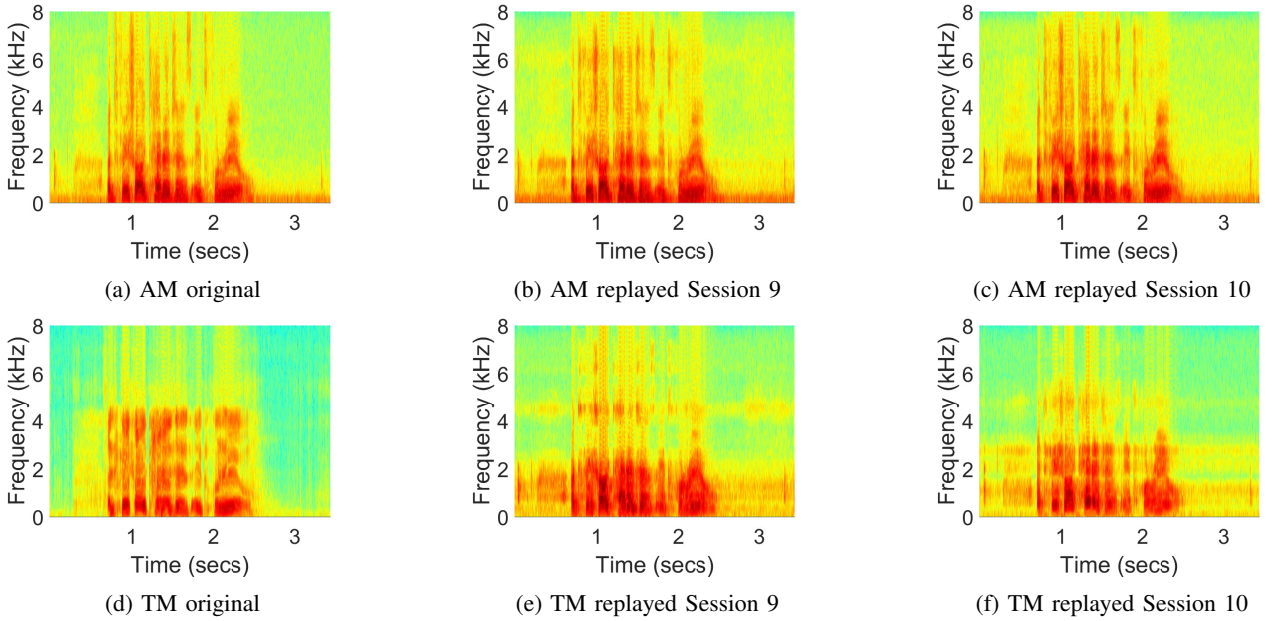
B. Collection of Replay Data

In commonly used designs to assess vulnerability of conventional single-microphone ASV systems against replay attack [14], [15], [10], the target speaker utterances are played back and re-recorded using a microphone. But the collection of replay data for the dual AM-TM system is slightly different: while the attacker can use the conventional replay method for the acoustic microphone, this is expected to be ineffective against the throat microphone which is a physical contact sensor and therefore the throat microphone would need to be kept very close to the replay loudspeaker. Thus, we instead physically attach the throat-mic to the attack loudspeaker to collect replayed throat microphone data, as illustrated in Figure 6 for three different replay configurations. Besides the conventional loudspeakers shown in the left and the middle photos, the photo on the right illustrates a replay configuration that uses a specific *audio exciter* to excite the throat microphone[5].

The test sections of the original acoustic mic data were played back and recorded in a different rooms using various loudspeakers. We have done playback recording in twelve different conditions. The details of the playback sessions are shown in Table I. We have total 300 genuine files in each

[4] http://us.focusrite.com/usb-audio-interfaces/scarlett-2i2

[5] An audio exciter is essentially a loudspeaker lacking a membrane; instead, it transmits the oscillations of audio signals to the surface where the exciter is physically attached. This causes the surface to vibrate and emit the audio signal. In our case, however the exciter is attached directly to the TM to directly stimulate the sensor inside the TM. It should be noted that due to the small surface area of the TM the sound produced by the exciter is not very loud and therefore does not interfere with the acoustic signal to the AM



(a) AM original (b) AM replayed Session 9 (c) AM replayed Session 10

(d) TM original (e) TM replayed Session 9 (f) TM replayed Session 10

Fig. 5: Spectrograms of the AM and TM signal of the passphase "Artificial intelligence is for real" from the original recording sessions, along with the replayed versions obtained in Session 9 and 10.

TABLE I: Description of setup for replay recordings.

Session	Room	Loudspeaker	Mic Distance
1	Room 1 (L=10m, W=8m, H=4m)	Logitech S-0264B	20cm
2	Room 2 (L=6, W=6m, H=2.5m)	Logitech S-0264B	15cm
3	Room 3 (L=3,5m, W=3,5m, H=2,5m)	JVC UX-BS1001	15cm
4	Room 4 (L=9m, W=3m, H=2.5m)	Logitech S-0264B	5cm
5	Room 5 (L=5m, W=4m, H=3m)	Creative A60	7cm, 50cm
6		GENELEC 8020C	16cm, 16cm
7		Logitech S-120	3cm, 24cm
8	Room 6 (L=20m, W=6m, H=3m)	Creative A60	6cm, 10cm
9	Room 7 (L=5,7m, W=4,2m, H=2,8)	JVC SP-EX70	8cm
10		VIFA M10MD-39-08	6cm
11		Eltex	5cm
12		VIFA M10MD-39-08 with HIHX14C02-8	4cm

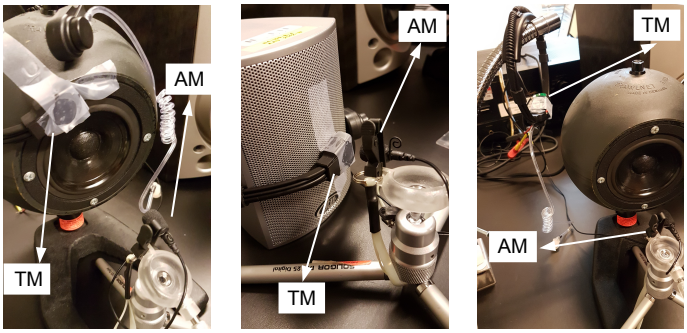


Fig. 6: Microphone setup for three replay conditions: Session 10 (left), Session 11 (middle) and Session 12 (right).

of the two test sessions. Hence we have total 3600 replay recording files by pooling all the replay sessions. Table II summarizes our database with the number of ASV trials for different conditions.

TABLE II: Trial summary for experimental evaluation.

Trial Type	Trial No
Genuine	300
Zero-effort Imposter	8700
Replay Imposter (300 from each of the 12 replay sessions)	3600

C. Description of Features and Classifiers for ASV

The recorded utterances were down-sampled from 44.1 kHz to 16 kHz to match the available off-line development data for training universal background model (UBM). We compute mel-frequency cepstral coefficients (MFCCs) as spectral features from both AM and TM signals. The MFCCs are extracted from speech frames of 20 ms duration, with 50% overlap. 20 filters in mel scale are used to compute 20 coefficients including the energy. Then RelAtive SpecTrAl (RASTA) processing [68] is used for suppressing linear channel effects. We obtain 60-dimensional features including

delta and double-delta coefficients computed with a window of three frames. Finally, we discard the non-speech frames using an energy-based SAD [1].

We consider two modern ASV systems, *Gaussian mixture model with universal background model* (GMM-UBM) [62] and an *i-vector* based system [63]. For both systems, the UBMs is gender-independent and trained from all 6300 speech files of the TIMIT corpus. TIMIT contains high-quality 16 kHz microphone speech in English similar in quality to our AM evaluation speech. We also use *domain adaptation* to create throat microphone UBM from TIMIT corpus. Domain adaptation, in general, refers to methodology for adapting the components of a well-trained recognizer (trained using large amount of out-of-domain data), to a new domain using a small amount of representative data from the new, or in-domain data [69]. In our case, we treat the acoustic mic data as out-of-domain data and the throat-mic data as the in-domain data. This choice is well-justified, since we are short of a large supply of throat-mic data but can first train the acoustic-mic UBM from a large dataset (here, TIMIT corpus of 630 speakers) and adapt it for the throat-mic data with only eight speakers. We also adapt the TIMIT UBM for limited acoustic microphone data also since adaptation for 10 pass-phrases could be beneficial. We train UBM of 512 Gaussians using 10 iterations of the expectation-maximization (EM) algorithm. The target speaker models are obtained using *maximum-a-posteriori* (MAP) adaptation of the Gaussian means with a relevance factor of 14. The i-vector extractor (*i.e.*, T-matrix) is trained for 400 total factors with five iterations of EM. We compute recognition score as log-likelihood ratio for GMM-UBM and cosine similarity for the i-vector system.

D. Performance Evaluation

We use *equal error rate* (EER) to assess both standalone liveness detection and ASV performance. EER corresponds to the detection threshold with equal false alarm rate (FAR) and false rejection rates (FRRs). We have used BOSARIS toolkit for computing the evaluation metrics with ROC Convex Hull method[6], a method suitable to obtain EER estimates from data with limited number of trials. We denote EERs for ASV and VLD systems as EER_{asv} and EER_{vld}, respectively. Besides EER, we also report false acceptance and false rejection rates as in [70] for the combined systems where ASV system is integrated with VLD method. We report two FARs separately computed for zero-effort impostor trials (no replay attack) and replay spoof trials. They will be referred to as FAR(Z) and FAR(R), respectively.

V. Results

A. Standalone ASV Performance

In the first experiment, we assess the accuracy of our ASV systems under both zero-effort imposter in matched (clean) and mismatched (noisy) condition. The results are shown in Table III for signals from individual microphones and fused system where scores of AM and TM are combined using equal weight fusion. In the third and the fourth columns, we have shown the results for the individual microphones as a baseline. Next, we have shown the results for domain adaptation where limited data of AM and TM are used to adapt the UBM trained with TIMIT.

The performance with the domain adaptation method shows improvement, noticeably for the throat microphone case. Then we exploit the fact TM signals are more robust than AM in presence of environmental noise and extract SAD labels from the TM speech, which is also used with the AM system. This gives improvement in recognition accuracy for AM, specially for mismatch condition. Further improvement is obtained by equal weight score fusion of AM and TM as reported in last column of Table III. For both the GMM-UBM and the i-vector systems, the AM-based system outperforms the TM-based system under matched condition, but the order is reversed in most cases under the noise mismatched condition. The details of the effect of SAD and fusion were reported in our preliminary study [27]. Our finding also agrees with another related prior work [25].

Next, we evaluate the performance of IRM-based weighting scheme as discussed in Section II-C. We experiment with different values of the tuning parameter, β, with the results shown in Table IV. The use of IRM-based enhanced signal helps in improving the performance over single channel based approaches. However, for the fused mode, *i.e.*, when combined with TM system, the performance of the proposed scheme where TM signal is used to compute the SAD labels and final score is computed by fusion of the AM and TM system, is still better. Note that the IRM-based method requires optimizing of β while the proposed dual-channel ASV method uses equal weights fusion without any additional parameters.

In order to validate whether the proposed scheme is suitable for other throat microphones, we have conducted ASV experiments with another dataset collected using a different throat-microphone model. To this end, we collected data from 13 additional speakers with more than six sessions in the test set compared with two sessions in our other dataset. The speakers were enrolled with clean speech data from three sessions similar to the previous case. The ASV performance with the GMM-UBM system is shown in Table V. As before, the relative degradation of TM-based ASV system in presence of mismatch, in comparison to that of the AM-based ASV system, is lower. Moreover, fusion helps to improve the overall recognition accuracy.

Next, we evaluate the ASV performance for the replay spoof condition. The results are shown in Table VI. Here, the scores of the original recording corresponding to the correct speakers are set as the target or genuine trials, and the scores of the replayed version of the same segments computed for their corresponding original source speakers, are set as the non-target or imposter trials. The results indicate that the performance is remarkably better for the throat-microphone compared to the acoustic microphone. For instance, in the matched condition, AM-based system yields EERs of 27.71% and 32.70% for the GMM-UBM and the i-vector systems, respectively, while the corresponding TM-based numbers are as low as 1.24% and 2.25%. The performance is also considerably improved with

[6] https://sites.google.com/site/bosaristoolkit/

TABLE III: Text-dependent ASV results (% of EER_{asv}) for **licit** condition. Results are shown for two separate microphones (AM and TM) for different setups with two different classifiers. We have also shown the performance of dual-mic system when equal weighted scores from AM (with TM for SAD) and TM based systems are combined.

ASV system	Condition	Baseline		Domain Adaptation		AM with	Fused
		AM	TM	AM	TM	TM-SAD	(AM+TM)
GMM-UBM	Matched	0.06	2.72	0.33	1.58	0.06	0.20
	Mismatched	10.33	6.67	8.67	4.40	4.40	1.44
i-vector	Matched	1.33	4.31	0.67	1.87	0.23	0.30
	Mismatched	12.67	9.00	9.72	4.38	7.10	1.71

TABLE IV: ASV performance using GMM-UBM system with IRM-based dual channel speech enhancement and and its score fusion with TM. The last row shows the results for proposed AM-based system using TM-based SAD and its fusion with TM-based system.

β	IRM-based enhanced AM		Fused with TM	
	Matched	Mismatched	Matched	Mismatched
1	1.76	4.58	1.72	4.39
2	0.22	1.96	0.48	2.27
3	0.12	1.89	0.29	1.84
4	0.12	2.32	0.23	1.73
5	0.10	2.59	0.21	1.60
6	0.10	3.17	0.20	1.59
7	0.10	3.41	0.21	1.54
8	0.10	3.62	0.22	1.50
9	0.10	3.73	0.22	1.49
10	0.10	3.81	0.22	1.49
Proposed	0.06	4.40	0.20	1.44

TABLE V: Text-dependent ASV results (% of EER_{asv}) for **licit** condition for a different dataset of 13 speakers collected with a different throat-microphone.

ASV system	Condition	AM	TM	Fused
GMM-UBM	Matched	0.55	5.47	0.94
	Mismatched	4.99	7.08	3.41

dual-mic based ASV system. However, for the spoof condition, TM-based system alone yields the lowest EER.

TABLE VI: Text-dependent ASV results for **spoof** condition. Results are shown in terms of EER (in %) for two separate microphones (AM and TM) using two different classifiers and their combined mode with equal weight fusion. SAD labels are always extracted with TM signals.

ASV system	Condition	AM	TM	Fused
GMM-UBM	Matched	27.71	1.24	6.78
	Mismatched	28.46	2.50	17.67
i-vector	Matched	32.70	2.25	5.34
	Mismatched	41.53	5.38	11.89

In another experiment, we varied the distance between the throat microphone and the replay loudspeaker to investigate which position of the throat microphone is the most difficult condition for the ASV systems. We keep the position of AM fixed. The EERs for clean, noisy and their averages are shown in Fig. 7 for four different distances: 0 cm (i.e., attached), 3 cm, 6 cm and 9 cm. We find that the scenario

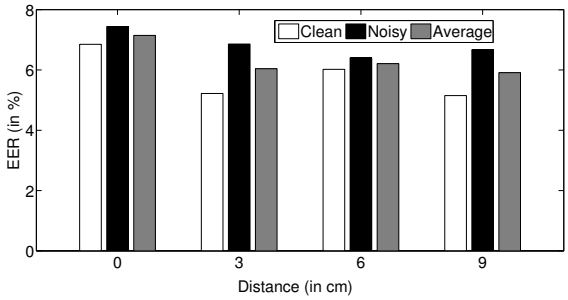


Fig. 7: Performance of proposed dual-mic ASV system (with GMM-UBM as a classifier) for different configurations of replay setup. Here the distance between the TM and the playback loudspeaker is varied but the position of AM is fixed.

when the TM is attached to the loudspeaker is relatively the most difficult condition. However, we have not observed any considerable difference by keeping the throat-microphone in a small distance apart.

B. Voice Liveness Detection Performance

In the replay spoofing scenario, the accuracy of standard ASV system is usually enhanced by integrating an auxiliary standalone countermeasure with the ASV system. Following a common evaluation strategy in the spoofing context [71], we first evaluate the accuracy of the countermeasures in isolation from any ASV system. To this end, we train our countermeasures with speech data from a set of eight speakers. Those eight speakers belong to two of the sites that participated in data collection, and this set is disjoint from the 30 speakers used in the evaluation set. The four clean sessions of these speakers are replayed and recorded in four different conditions. These four conditions are the last four replay sessions of Table I, i.e., Sessions 9 through 12. In total, we used 320 sentences to model the genuine class and 1280 sentences from total four replay sessions to model the spoof class. In test, we have total 300 genuine sentences and 3600 replay recordings from 12 sessions. Out of these 12 sessions, the replay settings for four of them (Sessions 9 through 12) are present in training data (referred to as *seen* condition) while the rest eight (Sessions 1 through 8) are absent in the training data (referred as *unseen* condition).

We separately assess the accuracy of MFCC, LFCC, and CQCC as the front-end spectral features with a GMM as the back-end. The MFCC features were extracted using 20

filters. 60-dimensional feature vector were formed by appending the delta and double-delta coefficients with the static coefficients. As for the CQCC and LFCC features, we extract 40-dimensional features by considering only the dynamic (*i.e.*, delta and double-delta) coefficients. We apply neither feature normalization nor speech activity detection. In the back-end, 512 Gaussians are used to model both live human and the replay classes.

TABLE VII: Standalone voice liveness detection performance (EER_{vld} for clean/noisy) for MFCC, LFCC and CQCC features using GMM-ML classifier with 512 mixtures. We have also shown the performance for combined dual-mic based system using equal weight score fusion and proposed joint feature.

Channel	System	Seen	Unseen	Pooled
AM	MFCC-60	3.31 / 16.14	19.97 / 37.84	15.67 / 30.96
TM	MFCC-60	0.00 / 0.00	6.83 / 3.27	5.55 / 2.57
Dual	Score Fusion	0.00 / 0.00	7.01 / 4.98	15.67 / 3.72
	Joint Feature	0.00 / 0.00	4.76 / 6.10	3.65 / 4.93
AM	LFCC-40	30.10 / 31.90	35.23 / 36.28	34.00 / 35.58
TM	LFCC-40	0.22 / 1.21	2.03 / 3.67	1.67 / 3.06
Dual	Score Fusion	0.27 / 1.68	1.84 / 2.96	1.46 / 2.64
	Joint Feature	0.00 / 0.00	0.00 / 0.04	0.00 / 0.03
AM	CQCC-40	21.58 / 24.92	30.53 / 31.40	27.53 / 29.67
TM	CQCC-40	5.10 / 9.71	7.33 / 12.83	6.83 / 12.00
Dual	Score Fusion	3.75 / 7.91	6.67 / 11.69	5.76 / 10.47
	Joint Feature	0.00 / 0.43	0.33 / 2.09	0.33 / 1.78

The performance of the stand-alone countermeasures is shown in Table VII for the seen and unseen conditions separately, as well as for the combined or 'pooled' case. Voice liveness detection accuracy is higher for the seen condition, as expected. The results further indicate that voice liveness detection performance of TM-based method systematically outperforms the AM-based approach by a large margin. This confirms our hypothesis that the TM signal is much harder to spoof.

The performance is also improved by combining the AM and the TM systems. Contrasting the score fusion and feature feature strategies, the latter is clearly more effective. The time-aligned joint distribution of AM and TM features distinguishes live and replay voices in a more efficient manner by capturing useful correlation across the two sensors. For this reason, the EERs are drastically reduced and sometimes even equal to zero[7]. We have also found that LFCCs are more efficient than baseline MFCC and state-of-the-art CQCC features.

C. Integration of ASV and Voice Liveness Detector

While the evaluation of the spoofing detector in isolation from an ASV system is instructive, the main question of interest is the accuracy of the final, integrated system. Integration of countermeasures with recognition system is an interesting open research problem within biometrics [72]. In this work, we adopt a simple *decision-fusion* method presented *e.g.* in [70], [73]. This is accomplished by 'AND'ing the decisions from ASV and VLD module which were computed with separate thresholds. For the ASV system, we set the detection threshold to the EER operating point using the 13 disjoint speakers that are not part of the evaluation. For the voice liveness detection system, in turn, we set the decision threshold to zero following the methodology used in [74].

[7]Similar to any corpus-based empirical speech processing research, this should be taken as an encouraging preliminary result, rather than a general claim of a solved problem, given the relatively small corpus size.

TABLE VIII: Speaker verification performance in terms of % of FAR and % of FRR for standalone and integrated system. Here, *Baseline* refers to the system with only acoustic mic. FAR(Z) and FAR(R) denote the false alarm rate for zero-effort imposter trials and replay imposter trials.

System	Condition	FAR(Z)	FAR(R)	FRR
Baseline ASV	Matched	0.10	84.25	0.00
	Mismatched	4.25	69.69	4.67
Baseline ASV with CM	Matched	0.10	18.17	10.33
	Mismatched	0.87	7.03	72.00
Dual-Mic ASV	Matched	0.03	39.67	0.67
	Mismatched	0.60	18.75	3.33
Dual-Mic ASV with CM	Matched	0.03	0.00	0.67
	Mismatched	0.60	0.00	4.33

We consider only the GMM-UBM based ASV system for this integration exercise, as shown in Table VIII. The baseline AM-based system is combined with the best AM-only countermeasure, *i.e.*, MFCC-60 shown in Table VII. On the other hand, the dual-mic system is combined with the best liveness detector that uses concatenated 80-dimensional LFCCs from AM and TM.

The results shown in Table VIII indicate that combining the countermeasure with ASV systems considerably helps in reducing the FAR(R). For example, FAR(R) drops from 84.25% to 18.17% for baseline AM-based system in matched condition. Similarly, for proposed dual-mic system, FAR(R) reduces from 39.67% (matched) and 18.75% (mismatched) to zero percent. Note that the performances in Table VIII are shown by pooling all trials from seen and unseen conditions.

VI. Conclusions

Conventional single-channel acoustic microphone ASV systems without any countermeasures are easily spoofed using replay attacks. Most of the existing study on replay attack detection report results of speech countermeasures with single microphone based systems. For improved protection of the ASV system from replay attacks, especially under adverse conditions, we proposed the use of an additional contact sensor, a throat microphone, to enhance ASV robustness *and* to enhance security under replay attacks. By its inherent nature, the throat microphone signal is difficult to spoof. Our experimental results agree well with our hypotheses. Specifically, we have found that (1) replay attack detection performance is considerably improved using throat microphones, (2) the joint use of information from two microphones also further helps in detecting replayed speech, consistently across clean and noisy conditions considered, and finally, (3) integrating the throat mic based robust countermeasure with the ASV system notably reduces the false acceptance rate.

Our work has a few limitations that should be addressed in future research. As there were no prior work or off-the-shelf corpus available to study replay attacks against throat-microphone based dual channel ASV systems, it was necessary to define the attack scenarios and collect a custom data consisting of a relatively low number of subjects (38). A larger scale testing is needed to claim generality of the results beyond our speaker population.

Nevertheless, advancing on most prior work on replay evaluations that almost exclusively used clean enviroments or homogenous rooms across training and test sets, we have demonstrated reasonable generality beyond training environments and -speakers. Our preliminary but encouraging findings clearly warrant further investigations to the general use of throat microphones, and potentially other non-acoustic sensors, in the ASV context under spoofing attacks. For instance, it would be interesting to study text-*independent* ASV scenarios, and to devise simplified energy- or correlation based countermeasures to benefit more effectively from the cross-channel information.

ACKNOWLEDGEMENTS

The authors would like to express their sincere thanks to the anonymous reviewers and the Associate Editor, whose critical remarks greatly influenced, and improved, the contents and presentation of our paper.

REFERENCES

[1] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Commun.*, vol. 52, no. 1, pp. 12 – 40, 2010.

[2] J. H. L. Hansen and T. Hasan, "Speaker recognition by machines and humans: A tutorial review," *IEEE Signal Process. Mag.*, vol. 32, no. 6, pp. 74–99, 2015.

[3] K. Lee, B. Ma, and H. Li, "Speaker verification makes its debut in smartphone," IEEE Signal Process. Soc. Speech and Language Technical Committee Newslett., 2013.

[4] Nuance, "Nuance vocalpassword," 2015. [Online]. Available: http://www.nuance.com/for-business/customer-service-solutions/voice-biometrics/vocalpassword/index.htm

[5] USAA, "Easily and securely log on to the USAA app," 2015. [Online]. Available: https://www.usaa.com/inet/pages/enterprise_howto_biometrics_landing_mkt?akredirect=true

[6] N. K. Ratha, J. H. Connell, and R. M. Bolle, "Enhancing security and privacy in biometrics-based authentication systems," *IBM Systems J.*, vol. 40, no. 3, pp. 614–634, 2001.

[7] Y. Lau, M. Wagner, and D. Tran, "Vulnerability of speaker verification to voice mimicking," in *Proc. Int. Symp. on Intell. Multimedia, Video and Speech Process.*, 2004, pp. 145–148.

[8] S. Prasad, Z.-H. Tan, and R. Prasad, "Multi-frame rate based multiple-model training for robust speaker identification of disguised voice," in *The 6th Int. Symp. on Wireless Personal Multimedia Commun. (WPMC)*. IEEE, 2013, pp. 1–4.

[9] J. Lindberg and M. Blomberg, "Vulnerability in speaker verification – a study of technical impostor techniques," in *Proc. European Conf. on Speech Commun. and Technol.*, Budapest, Hungary, September 1999.

[10] T. Kinnunen and et al., "Reddots replayed: A new replay spoofing attack corpus for text-dependent speaker verification research," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Process.*, New Orleans, USA, 2017.

[11] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: A survey," *Speech Commun.*, vol. 66, pp. 130–153, 2015.

[12] S. Ergunay, E. Khoury, A. Lazaridis, and S. Marcel, "On the vulnerability of speaker verification to realistic voice spoofing," in *Proc. IEEE 7th Int. Conf. Biometrics Theory, Applications and Systems (BTAS)*, Arlington, USA, Sept. 2015, pp. 1–6.

[13] F. Alegre, A. Janicki, and N. Evans, "Re-assessing the threat of replay spoofing attacks against automatic speaker verification," in *Proc. of the Int. Conf. of the Biometrics Special Interest Group*, Darmstadt, Germany, 2014, pp. 157–168.

[14] Z. Wu, S. Gao, E. Chng, and H. Li, "A study on replay attack and anti-spoofing for text-dependent speaker verification," in *Proc. Asia-Pacific Signal Inform. Process. Assoc. Annu. Summit and Conf.*, 2014, pp. 1–5.

[15] J. Galka, M. Grzywacz, and R. Samborski, "Playback attack detection for text-dependent speaker verification over telephone channels," *Speech Commun.*, vol. 67, pp. 143–153, 2015.

[16] T. Kinnunen and et al., "Utterance verification for text-dependent speaker recognition: a comparative assessment using the RedDots corpus," in *Proc. Conf. of the Int. Speech Commun. Assoc.*, San Francisco, USA, 2016, pp. 430–434.

[17] W. Shang and M. Stevenson, "Score normalization in playback attack detection," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Process.*, 2010, pp. 1678–1681.

[18] Z. Wu, T. Kinnunen, N. W. D. Evans, J. Yamagishi, C. Hanilçi, M. Sahidullah, and A. Sizov, "Asvspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," in *Proc. Conf. of the Int. Speech Commun. Assoc.*, Dresden, Germany, 2015, pp. 2037–2041.

[19] P. Korshunov and et al., "Overview of BTAS 2016 speaker anti-spoofing competition," in *Proc. Int. Conf. on Biometrics: Theory, Appl. and Syst.*

[20] M. Todisco, H. Delgado, and N. Evans, "A new feature for automatic speaker verification anti-spoofing: Constant Q cepstral coefficients," in *Proc. Odyssey: the Speaker and Language Recognition Workshop*, 2016.

[21] S. Shiota, F. Villavicencio, J. Yamagishi, N. Ono, I. Echizen, and T. Matsui, "Voice liveness detection for speaker verification based on a tandem single/double-channel pop noise detector," *Proc. Odyssey: the Speaker and Language Recognition Workshop*, pp. 259–263, 2016.

[22] S. Trewin, C. Swart, L. Koved, J. Martino, K. Singh, and S. Ben-David, "Biometric authentication on a mobile device: A study of user effort, error and task disruption," in *Proceedings of the 28th Annual Computer Security Applications Conference*, ser. ACSAC '12. New York, NY, USA: ACM, 2012, pp. 159–168. [Online]. Available: http://doi.acm.org/10.1145/2420950.2420976

[23] T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, and K. A. Lee, "The asvspoof 2017 challenge: Assessing the limits of replay spoofing attack detection," in *Proc. Interspeech 2017*, Stockholm, Sweden, 2017, pp. 2–6.

[24] M. Marx, G. Vinoth, A. Shahina, and A. Khan, "Throat microphone speech corpus for speaker recognition," *MES J. of Technol. and Manage.*, pp. 16–20, 2009.

[25] A. Shahina, B. Yegnanarayana, and M. Kesheorey, "Throat microphone signal for speaker recognition," in *Proc. Int. Conf. on Spoken Language Process.*, 2004.

[26] N. Mubeen, A. Shahina, A. Khan, and G. Vinoth, "Combining spectral features of standard and throat microphones for speaker identification," in *Proc. of Int. Conf. on Recent Trends In Inform. Technol.* IEEE, 2012, pp. 119–122.

[27] M. Sahidullah and et al., "Robust speaker recognition with combined use of acoustic and throat microphone speech," *Proc. Conf. of the Int. Speech Commun. Assoc.*, pp. 1720–1724, 2016.

[28] A. Jain, Y. Chen, and M. Demirkus, "Pores and ridges: High-resolution fingerprint matching using level 3 features," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 1, pp. 15–27, 2007.

[29] R. Derakhshani, S. Schuckers, L. Hornak, and L. O'Gorman, "Determination of vitality from a non-invasive biomedical measurement for use in fingerprint scanners," *Pattern Recognition*, vol. 36, no. 2, pp. 383–396, 2003.

[30] J. Galbally, F. Alonso-Fernandez, J. Fierrez, and J. Ortega-Garcia, "A high performance fingerprint liveness detection method based on quality related features," *Future Generation Computer Systems*, vol. 28, no. 1, pp. 311–321, 2012.

[31] P. Lapsley, J. Lee, D. Pare Jr, and N. Hoffman, "Anti-fraud biometric scanner that accurately detects blood flow," 1998, uS Patent 5,737,439.

[32] S. Patil and J. Hansen, "The physiological microphone (PMIC): A competitive alternative for speaker assessment in stress detection and speaker verification," *Speech Commun.*, vol. 52, no. 4, pp. 327 – 340, 2010.

[33] P. Korshunov and S. Marcel, "Cross-database evaluation of audio-based spoofing detection systems," in *Proc. Conf. of the Int. Speech Commun. Assoc.*, 2016.

[34] K. Patel, H. Han, and A. Jain, "Cross-database face antispoofing with robust feature representation," in *Chinese Conf. on Biometric Recognition*. Springer, 2016, pp. 611–619.

[35] C. Hanilçi, T. Kinnunen, M. Sahidullah, and A. Sizov, "Spoofing detection goes noisy: An analysis of synthetic speech detection in the presence of additive noise," *Speech Commun.*, vol. 85, pp. 83–97, December 2016.

[36] X. Tian, Z. Wu, X. Xiao, E. S. Chng, , and H. Li, "An investigation of spoofing speech detection under additive noise and reverberant conditions," in *Proc. Conf. of the Int. Speech Commun. Assoc.*, San Francisco, USA, 2016.

[37] W. Jin, S. Jou, and T. Schultz, "Whispering speaker identification," in *IEEE Int. Conf. on Multimedia and Expo*, 2007, pp. 1027–1030.

[38] A. Shahina and B. Yegnanarayana, "Language identification in noisy environments using throat microphone signals," in *Proc. of Int. Conf. on Intell. Sensing and Inform. Process.*, 2005, pp. 400–403.

[39] Z. Wu and et al., "ASVspoof: The automatic speaker verification spoofing and countermeasures challenge," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 4, pp. 588–604, 2017.

[40] Y. Li and D. Wang, "On the optimality of ideal binary time–frequency masks," *Speech Communication*, vol. 51, no. 3, pp. 230–239, 2009.

[41] S. Srinivasan, N. Roman, and D. Wang, "Binary and ratio time-frequency masks for robust speech recognition," *Speech Communication*, vol. 48, no. 11, pp. 1486–1501, 2006.

[42] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 22, no. 12, pp. 1849–1858, 2014.

[43] M. Kolbæk, Z.-H. Tan, and J. Jensen, "Speech intelligibility potential of general and specialized deep neural network based speech enhancement systems," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 153–167, 2017.

[44] J. Tardelli, "Pilot corpus for multisensor speech processing," Defense Technical Information Center, Tech. Rep., 2003.

[45] B. Denby, T. Schultz, K. Honda, T. Hueber, J. Gilbert, and J. Brumberg, "Silent speech interfaces," *Speech Commun.*, vol. 52, no. 4, pp. 270 – 287, 2010.

[46] S. Ballantine, "Contact microphone," July 1939, uS Patent 2,165,123. [Online]. Available: https://www.google.com/patents/US2165123

[47] "Iasus," http://iasus-concepts.com/, accessed: 2016-12-14.

[48] "Retevis," http://www.retevis.com/, accessed: 2016-12-14.

[49] "Astra radio communications," http://www.arcmics.com/professional-radio-accessories/throat-mics, accessed: 2016-12-14.

[50] Z. Zhang and et. al, "Multi-sensory microphones for robust speech detection, enhancement and recognition," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Process.*, vol. 3, 2004, pp. iii–781–4 vol.3.

[51] A. Stupakov, E. Hanusa, D. Vijaywargi, D. Fox, and J. Bilmes, "The design and collection of COSINE, a multi-microphone in situ speech corpus recorded in noisy environments," *Computer Speech & Language*, vol. 26, no. 1, pp. 52 – 66, 2012.

[52] M. Graciarena, H. Franco, K. Sonmez, and H. Bratt, "Combining standard and throat microphones for robust speech recognition," *IEEE Signal Process. Lett.*, vol. 10, no. 3, pp. 72–74, 2003.

[53] T. Dekens, Y. Patsis, W. Verhelst, F. Beaugendre, and F. Capman, "A multi-sensor speech database with applications towards robust speech processing in hostile environments." in *Proc. of the Int. Conf. on Language Resources and Evaluation*, 2008.

[54] T. Dekens, W. Verhelst, F. Capman, and F. Beaugendre, "Improved speech recognition in noisy environments by using a throat microphone for accurate voicing detection," in *Proc. European Signal Process. Conf.*, 2010, pp. 23–27.

[55] S. Jou, T. Schultz, and A. Waibel, "Adaptation for soft whisper recognition using a throat microphone." in *Proc. Conf. of the Int. Speech Commun. Assoc.*, 2004.

[56] V. Uloza, E. Padervinskis, I. Uloziene, V. Saferis, and A. Verikas, "Combined use of standard and throat microphones for measurement of acoustic voice parameters and voice categorization," *J. of Voice*, vol. 29, no. 5, pp. 552 – 559, 2015.

[57] F. Bozzoli and F. Angelo, "Measurement of active speech level inside cars using throat-activated microphone," in *Audio Eng. Soc. Conv. 116*. Audio Engineering Society, 2004.

[58] S. Ntalampiras, T. Ganchev, I. Potamitis, and N. Fakotakis, "Objective comparison of speech enhancement algorithms under real world conditions," in *Proc. of the 1st Int. Conf. on PErvasive Technol. Related to Assistive Environments*. ACM, 2008, p. 34.

[59] B. Acker-Mills, A. Houtsma, and W. Ahroon, "Speech intelligibility in noise using throat and acoustic microphones," *Aviation, space, and environmental medicine*, vol. 77, no. 1, pp. 26–31, 2006.

[60] M. Turan and E. Erzin, "Source and filter estimation for throat-microphone speech enhancement," *IEEE Trans. on Audio, Speech, and Language Process.*, vol. 24, no. 2, pp. 265–275, 2016.

[61] K. Vijayan and K. Murty, "Comparative study of spectral mapping techniques for enhancement of throat microphone speech," in *Twentieth Nat. Conf. on Commun. (NCC)*, 2014, pp. 1–5.

[62] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Process.*, vol. 10, no. 1-3, pp. 19–41, 2000.

[63] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. on Audio, Speech, and Language Process.*, vol. 19, no. 4, pp. 788–798, May 2011.

[64] M. Sahidullah, T. Kinnunen, and C. Hanilçi, "A comparison of features for synthetic speech detection," in *Proc. Conf. of the Int. Speech Commun. Assoc.*, Dresden, Germany, 2015, pp. 2087–2091.

[65] C. Hanilçi, T. Kinnunen, M. Sahidullah, and A. Sizov, "Classifiers for synthetic speech detection: A comparison," in *Proc. Conf. of the Int. Speech Commun. Assoc.*, Dresden, Germany, 2015, pp. 2057–2061.

[66] D. Reynolds and R. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 1, pp. 72–83, 1995.

[67] K. Lee and et al., "The RedDots data collection for speaker recognition," in *Proc. Conf. of the Int. Speech Commun. Assoc.*, 2015.

[68] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 578–589, Oct 1994.

[69] H. Daume III and D. Marcu, "Domain adaptation for statistical classifiers," *Journal of Artificial Intelligence Research*, vol. 26, pp. 101–126, 2006.

[70] M. Sahidullah, H. Delgado, M. Todisco, H. Yu, T. Kinnunen, N. Evans, and Z.-H. Tan, "Integrated spoofing countermeasures and automatic speaker verification: an evaluation on asvspoof 2015," *Proc. Conf. of the Int. Speech Commun. Assoc.*, pp. 1700–1704, 2016.

[71] P. Korshunov and S. Marcel, "Joint operation of voice biometrics and presentation attack detection," in *IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, 2016.

[72] I. Chingovska, A. Anjos, and S. Marcel, "Biometrics evaluation under spoofing attacks," *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 12, pp. 2264–2276, 2014.

[73] ——, "Anti-spoofing in action: Joint operation with a verification system," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2013.

[74] P. L. De Leon, M. Pucher, J. Yamagishi, I. Hernaez, and I. Saratxaga, "Evaluation of speaker verification security and detection of HMM-based synthetic speech," *IEEE Trans. on Audio, Speech, and Language Process.*, vol. 20, no. 8, pp. 2280–2290, 2012.

Md Sahidullah received the Ph.D. degree in the area of speech processing from the Department of Electronics and Electrical Communication Engineering of Indian Institute Technology Kharagpur in 2015. Prior to that he obtained the Bachelors of Engineering degree in Electronics and Communication Engineering from Vidyasagar University in 2004 and the Masters of Engineering degree in Computer Science and Engineering (with specialization in Embedded System) from West Bengal University of Technology in 2006. In 2007–2008, he was with Cognizant Technology Solutions India PVT Limited. Since 2014, he has been a post-doctoral researcher with the School of Computing, University of Eastern Finland. His research interest includes speaker recognition, voice activity detection and spoofing countermeasures.

Dennis A. L. Thomsen received the Master degree in the area of signal processing and computing from the Department of Electronic Systems of Aalborg University (AAU) in 2015. Since 2015, he has been a Research Assistant with AAU working on a H2020-funded OCTAVE project focusing on voice biometrics for physical and logical access control. His research interests include speech and speaker recognition, noise-robust speech processing, and voice activity detection.

Rosa Gonzalez Hautamäki received the M.Sc. degree in computer science from the University of Eastern Finland (formerly Univ. of Joensuu) in 2005, where she is currently pursuing a Ph.D. degree. Her research interests include speaker recognition and speech analysis.

Robert Parts received his MSc. degree in Automated Control Systems in 1981 from Tallinn Polytechnic Institute, Estonia. From 1981 to 1988 he worked as a programmer at the same institute and from 1988 to 1996 as a programmer at Mainor Parvi. Since 1996 to date, he has been a programmer at Adapt OÜ, Tallinn, Estinia, and since 2016 a co-owner of the same company. He is co-inventor in five granted patents concerning cloud computing security and authentication and a business partner in H2020-funded OCTAVE project.

Tomi Kinnunen received the Ph.D. degree in computer science from the University of Eastern Finland (UEF, formerly Univ. of Joensuu) in 2005. From 2005 to 2007, he was an associate scientist at the Institute for Infocomm Research (I2R) in Singapore. Since 2007, he has been with UEF. In 2010—2012, his research was funded by a post-doctoral grant from Academy of Finland focusing on speaker recognition. He was the PI in a 4-year Academy of Finland project focusing on speaker recognition and a co-PI of another Academy of Finland project focusing on audio-visual spoofing. He chaired *Odyssey 2014: The Speaker and Language Recognition workshop*. He served as an associate editor in *Digital Signal Processing* from 2013 to 2015. He currently serves as an associate editor in *IEEE/ACM Trans. on Audio, Speech and Language Processing* and *Speech Communication*. He is currently a partner in large H2020-funded OCTAVE project focusing on voice biometrics for physical and logical access control. In 2015–2016 he visited 6 months at National Institute of Informatics (NII), Japan, under a mobility grant from Academy of Finland, with focus on voice conversion, speaker verification and spoofing. He holds the honorary title of Docent at Aalto University, Finland, with specialization area in speaker and language recognition. He has authored and co-authored more than 100 peer-reviewed scientific publications in these topics.

Zheng-Hua Tan(M'00–SM'06) received the B.Sc. and M.Sc. degrees in electrical engineering from Hunan University, Changsha, China, in 1990 and 1996, respectively, and the Ph.D. degree in electronic engineering from Shanghai Jiao Tong University, Shanghai, China, in 1999. He is a Professor in the Department of Electronic Systems at Aalborg University, Aalborg, Denmark. He is also a co-head of the Centre for Acoustic Signal Processing Research (CASPR) at Aalborg University. He was a Visiting Scientist at the Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, USA, an Associate Professor in the Department of Electronic Engineering at Shanghai Jiao Tong University, and a postdoctoral fellow in the Department of Computer Science at Korea Advanced Institute of Science and Technology, Daejeon, Korea. His research interests include speech and speaker recognition, noise-robust speech processing, multimedia signal and information processing, human-robot interaction, and machine learning. He has authored or co-authored more than 170 publications in refereed journals and conference proceedings. He has served as an Editorial Board Member/Associate Editor for Elsevier Computer Speech and Language, Elsevier Digital Signal Processing, and Elsevier Computers and Electrical Engineering. He was a Lead Guest Editor of the IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING. He has served as a Chair, Program Co-chair, Area and Session Chair, and Tutorial Speaker of many international conferences.

Martti Pitkänen received the B.Sc.in Economics in 1973 from Helsinki School of Economics. Since then, he has working as a programmer, management accountant and production manager. Since 1982, he has been the owner and CEO of APLcomp Oy, Helsinki, Finland. He is a co-inventor in five granted patents concerning cloud computing security and authentication. He is a business partner in H2020-funded OCTAVE project.