di F. Rundo, S. Battiato, S. Conoci, A.L. Di Stallo

ANALISI VIBRAZIONALE AVANZATA (AVA): UN ALGORITMO INNOVATIVO PER L'ESECUZIONE DI INTERCETTAZIONI AMBIENTALI

Francesco RUNDO, ingegnere informatico, ha un dottorato di ricerca in Matematica Applicata conseguito presso l'Università di Catania. Svolge l'attività di R&D Engineer presso la STMicroelectronics, sviluppando algoritmi e modelli matematici per l'analisi dati in ambito industriale. Da anni svolge il ruolo di Consulente Tecnico di Parte nei contenziosi in ambito civile e penale per l'analisi matematica dei rapporti bancari e degli strumenti di investimento.

Sebastiano BATTIATO è professore ordinario in Informatica presso l'Università di Catania. Docente dei corsi di Multimedia, Computer Vision e Computer Forensics. Consulente tecnico della Procura della Repubblica presso il Tribunale di Milano, Roma, Napoli, Bergamo e Reggio Calabria per attività di "image and video forensics".

Sabrina CONOCI, laureata in Chimica Industriale presso l'Università degli Studi di Bologna e con dottorato di ricerca in Ingegneria dei Materiali, ricopre il ruolo di R&D Manager presso STMicroelectronics occupandosi si attività di ricerca nell'ambito del settore dei dispositivi nanomolecolari, biosensori per l'analisi del DNA e sensori chimici.

Agatino Luigi DI STALLO, avvocato, è socio fondatore dello studio legale Di Stallo&Partners nonché componente del Direttivo Nazionale del network giuridico Master Legal Service(MLS). È co-fondatore del Laboratorio Scientifico-Giuridico "Giurimatica" dedicato allo studio teorico ed all'applicazione pratica delle scienze matematiche al diritto.

Decreto legislativo 29 dicembre 2017, n. 216

Con decorrenza Gennaio 2018, entra in vigore la riforma della disciplina delle intercettazioni attuata con il decreto legislativo 29 dicembre 2017, n. 216: "*Disposizioni in materia di intercettazioni di conversazioni o comunicazioni, in attuazione della delega di cui all'articolo 1, commi 82, 83 e 84, lettere a), b), c), d) ed e), della legge 23 giugno 2017, n. 103*".

La riforma rafforza il ruolo delle intercettazioni come indispensabile strumento di indagine ed investigazione forense. Nell'ambito di questa riforma, gli autori intendono illustrare nel presente articolo, una innovativa pipeline di elaborazione dati che consente, sotto opportune ipotesi, *la ricostruzione delle conversazioni tra due o piu' soggetti, dal solo filmato video (a risoluzione e frame-rate tipici di uno dispositivo di acquisizione video di ultima generazione) senza avere accesso alla relativa traccia audio associata.*

1.0 Panoramica sulla normativa in merito alle intercettazioni ambientali. Criticità dei sistemi attuali

Recentemente è stata approvata la riforma della disciplina delle intercettazioni attuata con il decreto legislativo 29 dicembre 2017, n. 216. Il decreto, recante "***Disposizioni in materia di intercettazioni di conversazioni o comunicazioni, in attuazione della delega di cui all'articolo 1, commi 82, 83 e 84, lettere a), b), c), d) ed e), della legge 23 giugno 2017, n. 103***", è stato pubblicato sulla Gazzetta Ufficiale n. 8 dell'11 gennaio scorso. La riforma conferma il ruolo delle intercettazioni come fondamentale strumento di indagine

e mira a creare un giusto equilibrio tra la segretezza della corrispondenza e di ogni altra forma di comunicazione e il diritto all'informazione. Pertanto, alla luce del suddetto riassetto normativo, appare certamente idoneo discutere di nuovi approcci per l'esecuzione delle intercettazioni, nello specifico, quelle ambientali, considerate peraltro le nuove disposizioni in tema di utilizzo dei cosiddetti "***trojan horse***" .
Le intercettazioni ambientali sono realizzate in una moltitudine di modi, impiegando le più svariate tecnologie. Prevalentemente vengono utilizzati *microspie, microregistratori, microfoni direzionali, tracciamento GPS, ecc.* Ad ogni modo il posizionamento di questi dispositivi è operazione rischiosa e complessa che può in alcune ipotesi rallentare o addirittura vanificare le attività investigative, soprattutto quando le controparti intercettate, ipotizzandone la presenza, operano opportune *bonifiche ambientali* volte a vanificare l'operato degli inquirenti. Dunque, una tecnologia che permetta di acquisire una conversazione tra due soggetti (ed il relativo filmato video), senza il posizionamento di microspie o altri captatori elettronici equivalenti, può certamente migliorare il lavoro degli investigatori, in questa delicata fase di indagine.

2.0 AVA: Analisi Vibrazionale Avanzata per l'esecuzione di intercettazioni ambientali

Quando un'onda sonora (vibrazione meccanica) investe un oggetto, genera delle micro-vibrazioni superficiali nell'oggetto stesso, spesso impercettibili all'occhio umano. Un team di ricercatori del prestigioso MIT di Boston, guidati dal Prof. Freeman, ha messo a punto un sistema algoritmico-matematico che analizzando esclusivamente informazioni video ad alto frame-rate, dell'oggetto investito dall'onda sonora, è in grado di estrarre *micro-vibrazioni* e recuperare con buona approssimazione il suono che originariamente, le ha prodotte. Tutte le informazioni su tale ricerca si trovano in [1]. Questo algoritmo, pertanto, è in grado di ripristinare i suoni da filmati "high speed" riferiti ad una varietà di oggetti con proprietà fisiche-chimiche diverse.

In questo articolo gli autori, tuttavia, propongono una piccola modifica all'algoritmo originario, integrando recenti tecniche di *machine learning*. Il metodo di seguito illustrato, rileva visivamente piccole vibrazioni in un oggetto, che rispondono al suono che lo sta investendo superficialmente, convertendo queste vibrazioni in un segnale audio verosimilmente prossimo a quello originario, trasformando pertanto, oggetti visibili di tutti i giorni in potenziali microfoni. Per recuperare il suono dalle micro-vibrazioni prodotte sull'oggetto a cui è diretto, preliminarmente è necessario acquisire un video senza audio, anche a distanza, in cui è presente l'oggetto ed il suono da recuperare, utilizzando una videocamera con alto frame-rate. Oggigiorno, i dispositivi video di ultima generazione sono dotati di sensori di acquisizione capaci di registrare video con elevata risoluzione e ad alto frame-rate, per cui l'algoritmo proposto può essere riprodotto anche a costi sostenibili. Una volta ottenuto il video, le informazioni di movimento vengono estratte attraverso una tecnica nota come "*Complex Steerable Pyramid*" descritta in [2].

Questi segnali di moto (micro-vibrazioni opportunamente processate) saranno dunque, *allineate e mediate* ulteriormente in un singolo segnale di movimento 1D (monodimensionale) che è correlato al movimento globale dell'oggetto nel tempo. Alla fine, questo segnale sarà ulteriormente filtrato in frequenza e depurato dal rumore, al fine di ricostruire il suono originario correlato alle informazioni di movimento (micro-vibrazioni). La ricostruzione del segnale di partenza dai segnali di moto, oltre che dal metodo sopra esposto, viene parallelamente ricostruita da un sistema di "**Stacked Autoencoders(SAEs)"** con *Softmax* layer di output, opportunamente addestrato, che si occuperà di ricostruire il suono originario mediante algoritmi di apprendimento di tipo "*Error Back Propagation*" con "*Regularized Empirical Risk*" al fine di evitare problematiche di *overfitting* del sistema [3]. Questo metodo, proposto dagli autori, rende piu' accurato e robusto il procedimento di ricostruzione e stima del suono di partenza, rispetto al metodo originariamente proposto dal team dell'MIT. Si osservi che la propagazione delle onde sonore in un materiale dipende da vari fattori: *la densità del materiale, la comprimibilità del materiale, la forma*, ecc.. In [1]-[2] è riportato uno studio ampio e particolareggiato in cui si analizzano le risposte analitiche di diversi oggetti e materiali. La seguente figura mostra la pipeline che gli autori propongono per la ricostruzione del segnale audio:

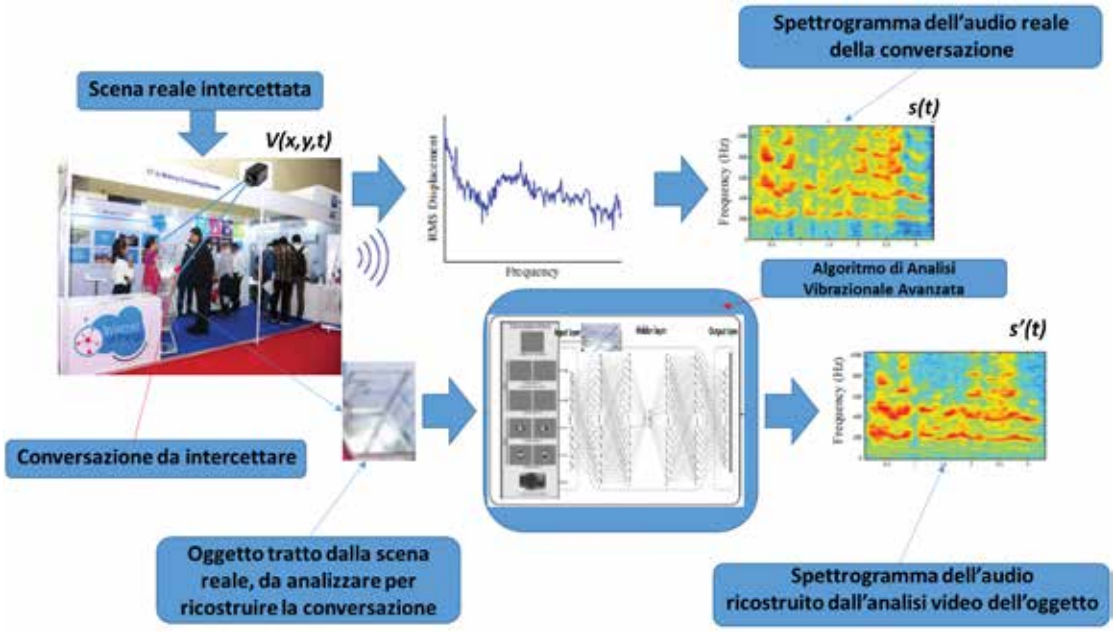

Figura 1 - *La pipeline proposta per l'esecuzione di intercettazioni ambientali*

Di seguito alcuni richiami matematici dell'algoritmo sul quale si basa la piattaforma AVA (Analisi Vibrazionale Avanzata) che viene proposta nel presente contributo. Il suono in ingresso ***s(t)*** (il segnale che vogliamo recuperare) è composto da fluttuazioni o onde meccaniche che esercitano una certa pressione sulla superficie di alcuni oggetti, nello specifico, riferendoci alla Figura 1, ci concetriamo, ad esempio, all'oggetto in plexiglass che si trova tra i due soggetti che stanno conversando e la cui conversazione (segnale **s(t)**) vorremmo ricostruire dal solo segnale video **V(x,y,t),** ripreso dal sistema di acquisizione posto in alto nella scena.

Le suddette fluttuazioni o onde meccaniche, generano delle micro-vibrazioni nell'oggetto (impercettibili ad occhio nudo), determinando una dinamica vibrazionale nel tempo che sebbene apparentemente invisibile, è di fatto presente e dunque lo sarà anche nel filmato video che stiamo acquisendo. Nello scenario di figura 1,ad esempio, le micro-vibrazioni generate dalla conversazione dei sue soggetti, nell'oggetto in plexiglass che si trova tra di loro, saranno incluse nel video che sta filmando l'intera dinamica reale. A questo punto, elaboriamo il video registrato con il nostro algoritmo per ricostruire il suono ***s'(t)***.
Dunque, schematizzando la pipeline: l'input del nostro sistema è un video **V(x,y,t)** di un oggetto**,** acquisito tipicamente con un frame-rate nel range 1kHz-20kHz. Assumiamo che in questo video sono incluse implicitamente le micro-vibrazioni legate ad un suono **s(t)** che ha investito la superficie di questo oggetto. Il nostro obiettivo è ricostruire **s(t**) da **V(x,y,t)** stimandolo con un segnale ricostruito **s'(t).**

Procediamo in tre passi. Per prima cosa, scomponiamo il video di input **V(x,y,t)** in sottobande tempo-spaziali corrispondenti a diversi orientamenti θ e scale ***r***. Quindi calcoliamo i segnali di *motion* locali e per un sotto-insieme di pixel, di orientamento e di scala. Combiniamo questi segnali di motion attraverso una sequenza di operazioni di media e allineamento per produrre un singolo segnale 1D correlato al movimento globale dell'oggetto. Parallelamente, i suddetti segnali di moto, andranno in input ad un sistema di **SAEs** opportunamente addestrato, il quale ricostruira il segnale **s'(t).** Infine, applichiamo delle tecniche classiche di *denoising* audio e tecniche di filtraggio in frequenza del segnale, per migliorarne la qualità.

Per calcolare, nel primo passo, le sottobande spaziali corrispondenti a differenti orientamenti e scale, applichiamo il metodo *Complex Steerable Pyramid* [2] usando funzioni-basi ***Gabor-like wavelets***. Per il filtraggio e la separazione di ampiezza e fase, nella trasformazione wavelet, usiamo filtri seno e coseno. Pertanto, ad ogni scala ***r*** ed orientamento θ possiamo associare un'immagine complessa che può essere espressa in termini di ampiezza **A** e fase φ**:**

$$A(r,\theta,x,y,t) = e^{i\varphi(r,\theta,x,y,t)}$$

Ciò premesso, prendiamo i valori di fase locali φ come sopra calcolabili e vi sottraiamo il valore della fase locale riferita ad un frame di riferimento t_0 (in genere il primo frame del video) per calcolare le variazioni assolute di fase:

$$\varphi_v(r,\theta,x,y,t) = \varphi(r,\theta,x,y,t) - \varphi(r,\theta,x,y,t_0)$$

È stato dimostrato che in ipotesi di micro-movimenti (o movimenti infinitesimali), queste variazioni di fase sono approssimativamente proporzionali a micro-spostamenti nella struttura dell'immagine sorgente lungo la direzione corrispondente l'orientamento θ e la scala **r** [4].
Nel metodo *Complex Steerable Pyramid*, per ogni orientamento θ e scala **r**, calcoliamo una media spaziale-ponderata dei segnali di *motion* locali al fine di ottenere un unico segnale di movimento Φ**(r,θ,t)** per ciascuna coppia (θ, r). Questa operazione ci permette di compensare alcuni *drawbacks* del suddetto approccio[2], riscontrabili principalmente nelle aree di immagini dove non vi è sufficiente contenuto informativo (textures) il che produrrebbe, inevitabilmente molto rumore nel calcolo delle variazioni di fase, come da equazioni sopra riportate. Applicando il suddetto approccio, la variazione di fase unica per ciascun *motion signal* i-esimo, può essere calcolata facendo riferimento ad una misura media dell'ampiezza **A** (correlata al contenuto informativo della relativa texture), tipicamente, il suo valore quadratico pesato opportunamente dalla corrispondente variazione di fase assoluta(rispetto al frame t_0):

$$\Phi_i(r,\theta,t) = \sum_{x,y} A(r,\theta,x,y,t)^2\, \varphi_v(r,\theta,x,y,t)$$

Prima di calcolare la media di Φ**(r,θ,t)** su scale e orientamenti diversi, gli autori in [2] suggeriscono un allineamento temporale per evitare distorsioni nel segnale risultante dalla media. Pertanto, il segnale unico ottenuto da questo allineamento temporale e dalla media pesata lungo ciascun orientamento e scala i-esimo, potrà essere calcolato risolvendo un problema di ottimizzazione come di seguito esposto:

$$\Phi(r_i,\theta_i,t-t_i) \;:\; t_i = arg\min_{t_i} \Phi_0(r_0,\theta_0,t)^T \Phi_i(r_0,\theta_0,t)$$

Dove con Φ_0**(r**$_0$**,θ**$_0$**,t)** si rappresenta un segnale *motion* di riferimento, scelto arbitrariamente durante il procedimento di *averaging*(media). A questo punto, il nostro segnale audio stimato **s'(t)** correlato al segnale unico di *motion*, come sopra calcolato, sarà cosi determinato (opportunamente normalizzato):

$$s'(t) = \sum_i \Phi(r_i,\theta_i,t-t_i)$$

Gli autori suggeriscono, altresì, di abbinare in parallelo, un apprendimento di tipo "Error Back Propagation" sui dati $\Phi_i(*)$ con target **s(t)** mediante l'utilizzo di sistemi SAEs con un layes neurale *hidden* ed uno strato di uscita *Softmax* per la ricostruzione del segnale. Il segnale cosi stimato sarà denominato s_n'**(t)** e potra essere opportunamente mediato con **s'(t)** al fine di migliorarne la qualità. Infine, il segnale audio stimato **s'(t**) sarà ulteriormente processato al fine di migliorare il Signal to Noise ratio(SNR) ovvero il rapporto segnale-rumore. In [2] i ricercatori che hanno validato estesamente questa pipeline, hanno notato la presenza

di una intensa componente di rumore alle basse frequenze, dell'audio ricostruito. Per risolvere questo problema, molti autori suggeriscono l'utilizzo di un filtro Butterworth passa-alto con frequenza di cutoff nel range 20-100Hz o in generale, pari ad 1/20 della frequenza di Nyquist, del segnale ricostruito. Talvolta, questo filtraggio è stato applicato al segnale $\mathbf{\Phi}(*)$, prima del procedimento di media ed allineamento sopra descritto. Una possibile limitazione della tecnica presentata finora è correlata alla necessità di acquisire un video con un frame-rate ***alto***, sebbene l'hardware oggi disponibile permette di superare agevolmente questo *constraints*, a costi peraltro sostenuti. Ad ogni modo in [2] gli autori suggeriscono diverse tecniche per eliminare questo limite. L'esecuzione delle suddette post-elaborazioni può essere eseguita su un framework che gli autori propongono composto da microcontrollori a 32 bits [5] con elevate capacità computazionali che si potranno occupare di eseguire l'algoritmo *Complex Steerable Pyramid* ovvero il pre-processing temporale ed il post-processing e filtraggio in frequenza , sopra descritto.

Per la parte relativa all'apprendimento SAEs, gli autori propongono l'utilizzo di sistemi hardware ad-hoc capaci di eseguire il learning previsto per i suddetti modelli. I sistemi descritti in [6] ben si addicono a questo genere di computazioni in quanto presentato un ottimo trade-off tra capacità di calcolo e velocità di esecuzione. Nel video riportato in [7] e nella pagina descrittiva del team di ricercatori dell'MIT [1], sono riportati degli esempi applicativi che mostrano la validità del metodo ivi illustrato.

3.0 Conclusioni e visione prospettica

Nel presente contributo gli autori hanno mostrato come la piattaforma AVA basata sull'analisi delle micro-vibrazioni di oggetti, di uso comune, in risposta al suono (onda meccanica) che li investe in superficie, può essere usata efficacemente per recuperare il relativo audio, trasformando pertanto questi oggetti in "microfoni visivi". L'ulteriore layer di ricostruzione intelligente, proposta dagli autori e basata su sistemi SAEs rende più efficace il metodo proposto sicchè quest'ultimo può essere valutato come valido strumento per l'esecuzione di intercettazioni ambientali a costi sostenibili e con metodiche di altissimo spessore scientifico. Ovviamente, esistono dei punti su cui la suddetta tecnologia può essere ulteriormente migliorata, tuttavia rimane un buon punto di partenza per innovare questo settore di cruciale importanza nell'indagine forense.

Bibliografia

[1] http://people.csail.mit.edu/mrub/vidmag/
[2] Abe Davis, Michael Rubinstein, Neal Wadhwa, Gautham J. Mysore, Fredo Durand, William T. Freeman, "The Visual Microphone: Passive Recovery of Sound from Video" , ACM Transactions on Graphics (Proc. SIGGRAPH), 2014, Vol. 33(4) pag. 79:1--79:10.
[3] Bishop, C. M. Pattern Recognition and Machine Learning. Springer, New York, NY, 2006.
[4] Gautama, t., and Van Hulle, "A phase-based approach to the estimation of the optical flow field using spatial filtering". IEEE Transactions on Neural Networks, 2002, 13, 5 (sep), 1127 – 1136;
[5] http://www.st.com/en/microcontrollers/stm32-32-bit-arm-cortex-mcus.html , 2018
[6] http://www.st.com/content/st_com/en/about/media-center/press-item.html/t4010.html , 2018
[7] https://www.youtube.com/watch?v=FKXOucXB4a8. ©