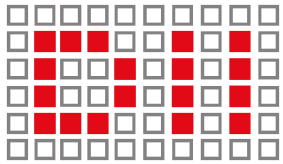
Scuola di Dottorato in Ingegneria dell'Informazione

XXXI Ciclo n.s., 3° anno di corso (2017/2018)

Studente:
Paolo Vecchiotti

Università Politecnica delle Marche

Attività di Ricerca



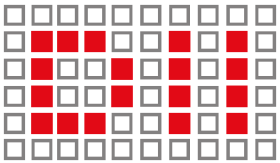
Tematiche trattate:

- Voice Activity Detection (VAD)
- Speaker Localization (SLOC)
- Integrazione di VAD e SLOC

Altri contributi:

- Tecniche di filtraggio avanzate per crossover audio

Ricerca: Overview



Elementi fondamentali:

- VAD e SLOC: fondamentali nello speech processing
- Interesse per l'ambiente smart-home
- Utilizzo di Reti Neurali

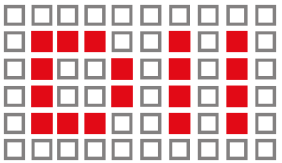
Motivazioni:

- Reti neurali più affidabili di algoritmi classici
- Algoritmi classici richiedono un fine tuning oneroso
- Gli algoritmi classici fanno difficoltà a generalizzare
- Risultanti promettenti ottenuti in letteratura con reti neurali applicate all'audio

Problematiche:

- Riverbero
- Cross-talk
- Rumore

Ricerca : VAD



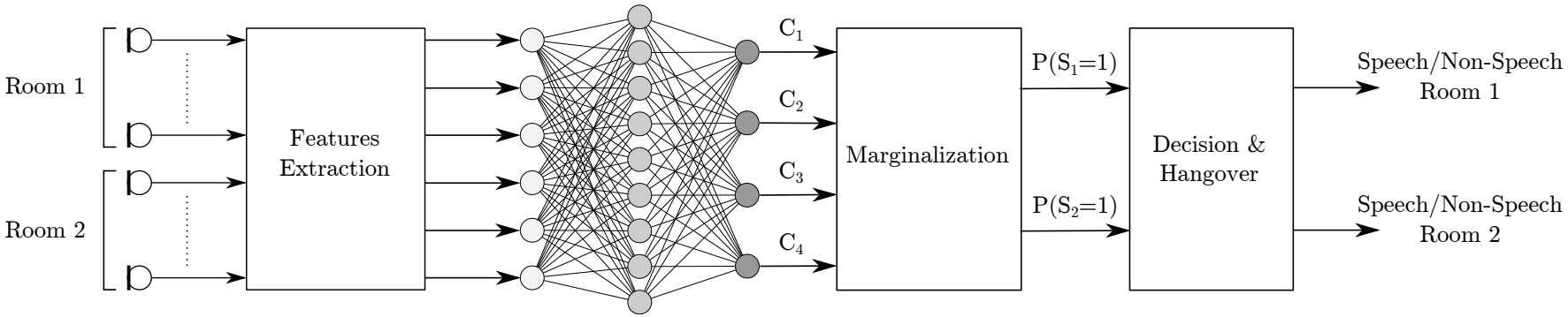
Elementi chiave:

- Determinare quando un soggetto umano sta parlando
- Classificazione attuata da reti neurali
- Ambiente multi-room
- Possibilità di sfruttare più microfoni
- Features estratte dal segnale registrato
- Utilizzare in maniera cooperativa il segnale proveniente da più stanze

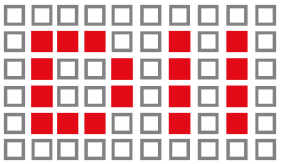
Ricerca : VAD

Studio Preliminare

- Algoritmo composto da più stage
  - Features Extraction
  - Neural Network
  - Marginalization
  - Decision&Hangover



Ricerca : VAD



Studio Preliminare

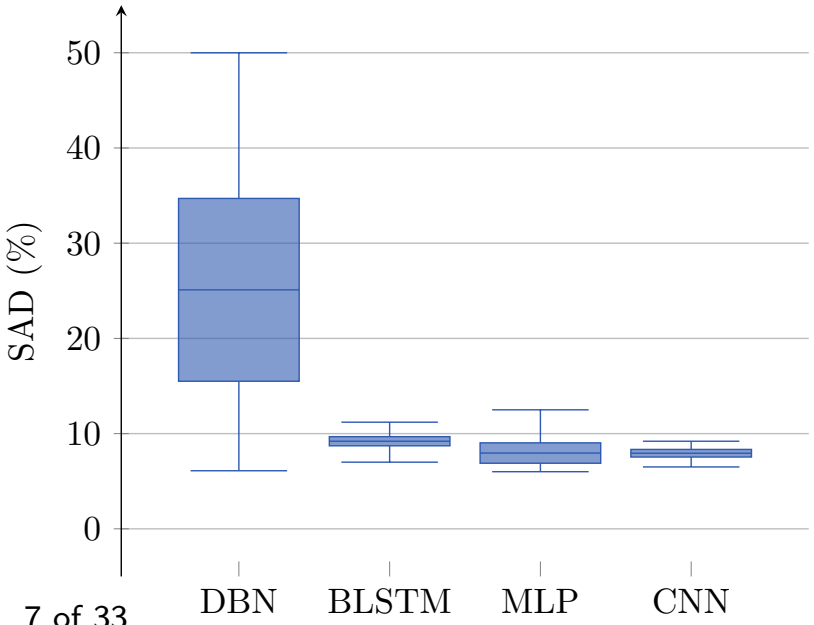
- Features Utilizzate
  - Envelope-Variance Measure
  - Pitch
  - WC-LPE Feature
  - Mel-Frequency Cepstral Coefficient
  - RASTA-PLP
  - Amplitude Modulation Spectrum
  - LogMel
- Reti Neurali Utilizzate
  - Multilayer Perceptron
  - Deep Belief Network
  - Bidirectional Long-Short Memory
  - Convolutional Neural Network

Ricerca : VAD

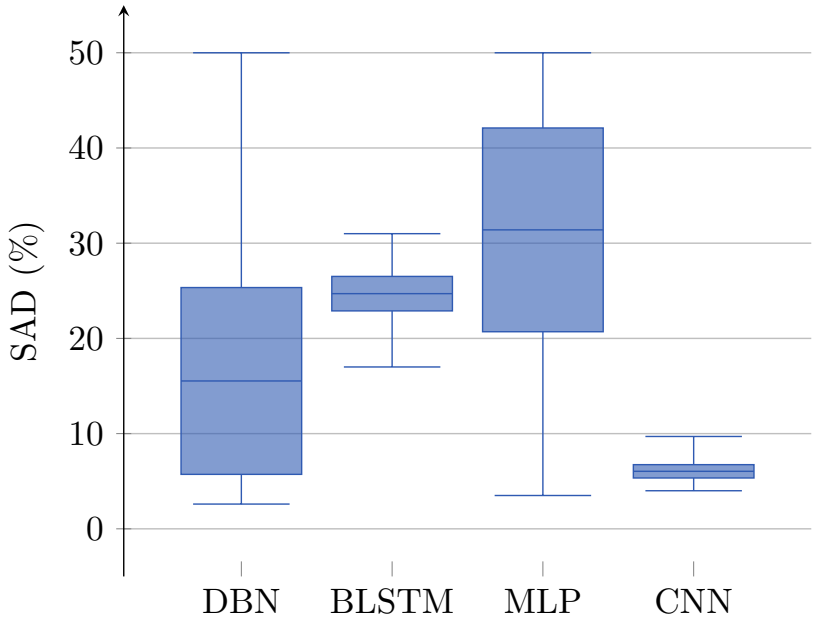
Studio Prelininare - Risultati

- Errore in SAD (%)
- Performance migliori in assoluto: DBN
- CNN dimostrano maggiore robustezza al posizionamento dei microfoni

DIRHA Simulated



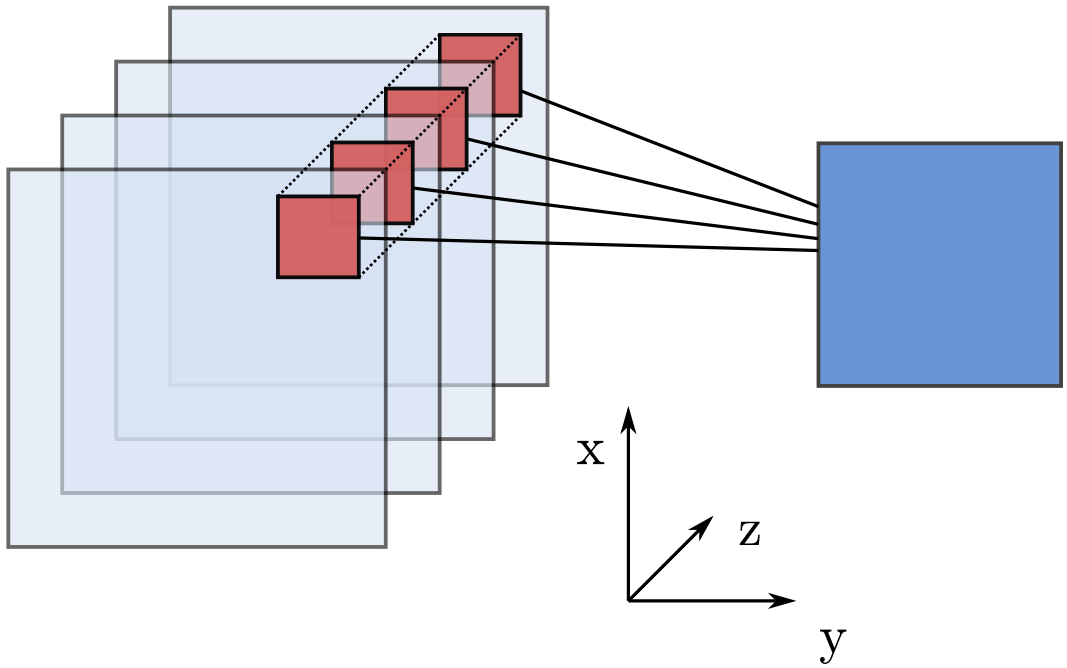
DIRHA Real



Ricerca : VAD

Advancements

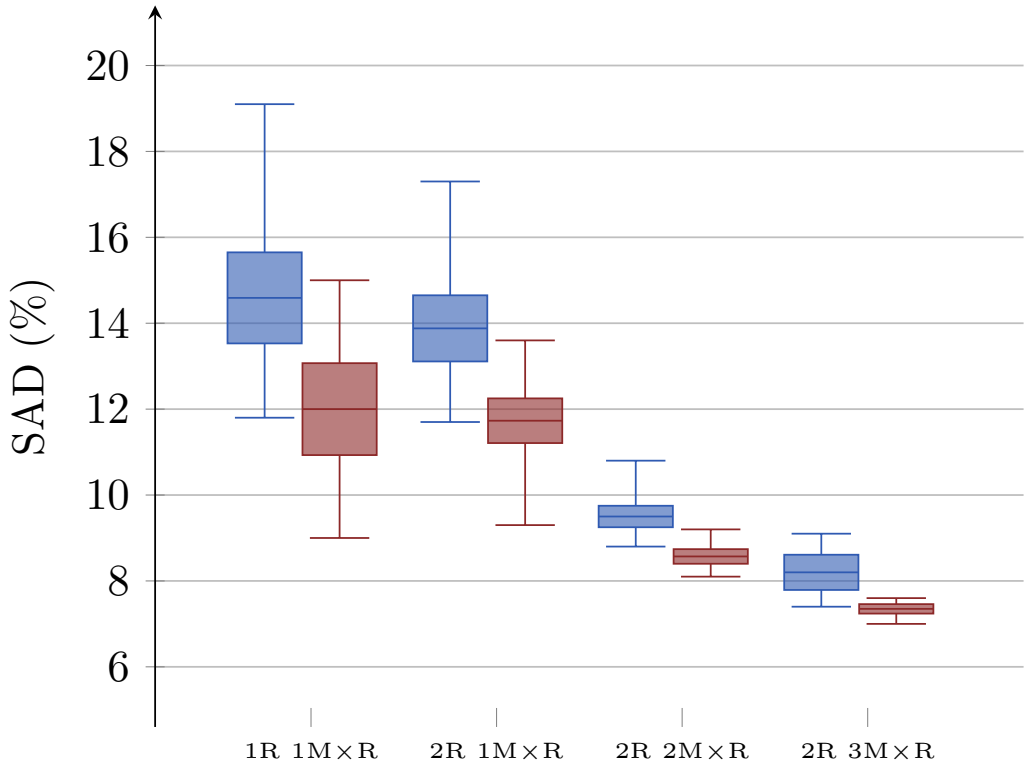
- Interesse verso nuove tecniche
  - Reti Neurali Convoluzionali con Kernel 3-D
  - Sistemi multi-channel che utilizzano più microfoni
  - Utilizzare simultaneamente dati provenienti da più stanze



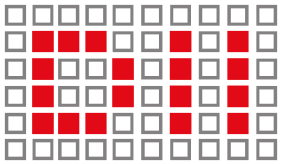
Ricerca : VAD

Advancements - Risultati

- CNN (rosso) migliore di MLP (blu)
- l'utilizzo di più microfoni comporta prestazioni migliori



Ricerca : SLOC - Approccio I



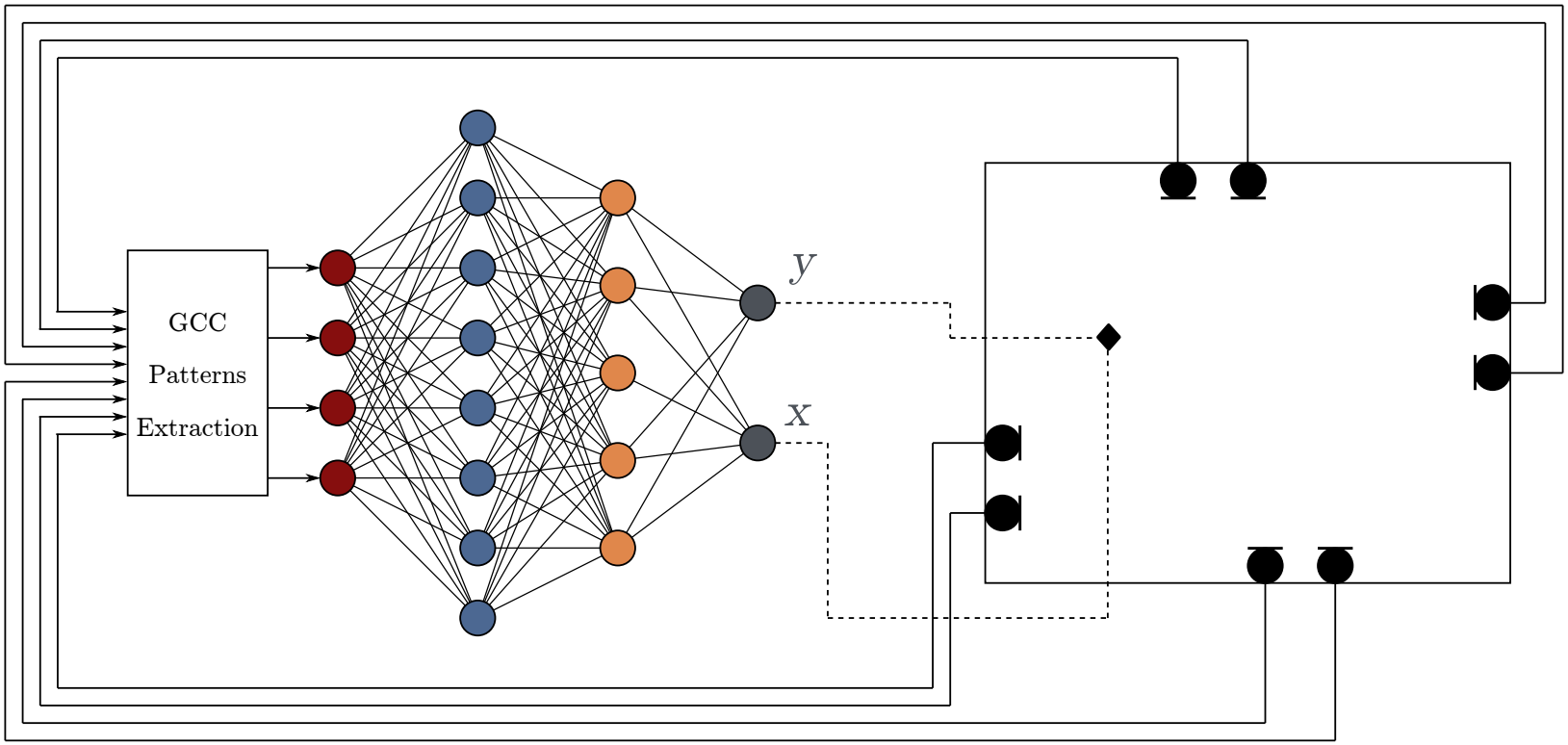
Elementi chiave:

- Determinare la posizione del parlatore nella stanza
- Utilizzo di Reti Neurali
- Localizzazione puntuale al posto della localizzazione della Direzione di Arrivo (DOA)
- Primo modello in letteratura

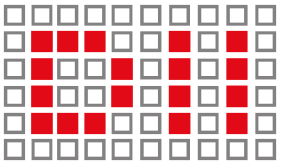
Ricerca : SLOC - Approccio I

Studio Preliminare

- Algoritmo composto da due stage
  - Features Extraction: GCC-PHAT Patterns
  - Neural Network: Multi-Layer Perceptron (MLP)



Ricerca : SLOC - Approccio I



Studio Preliminare

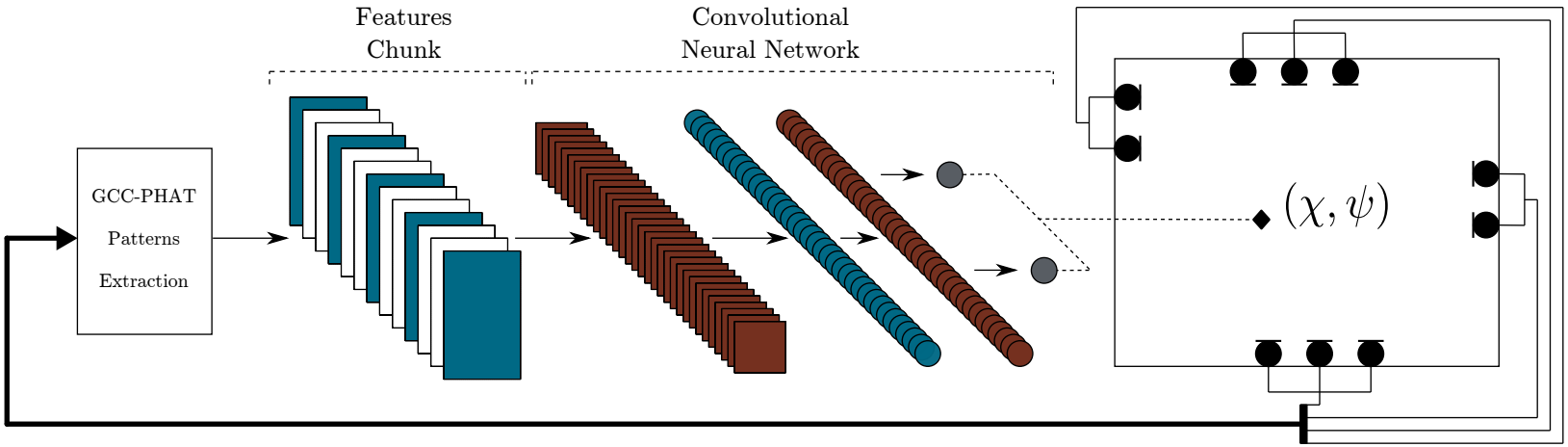
- Paragonato con algoritmo Cross Spectrum Phase (CSP-SLOC), algoritmo non neurale comunemente usato
- Errore misurato in Root Mean Square Error (RMSE)

Room	Algorithm	RMSE (mm)
Kitchen	CSP-SLOC	1280
	Proposed	**475**
Living Room	CSP-SLOC	1650
	Proposed	**525**

Ricerca : SLOC - Approccio I

Advancements

- Convolutional Neural Networks e Multi Layer Perceptron
- CNN: Utilizzo parallelo di più microfoni non possibile con MLP
- Studio del contesto temporale



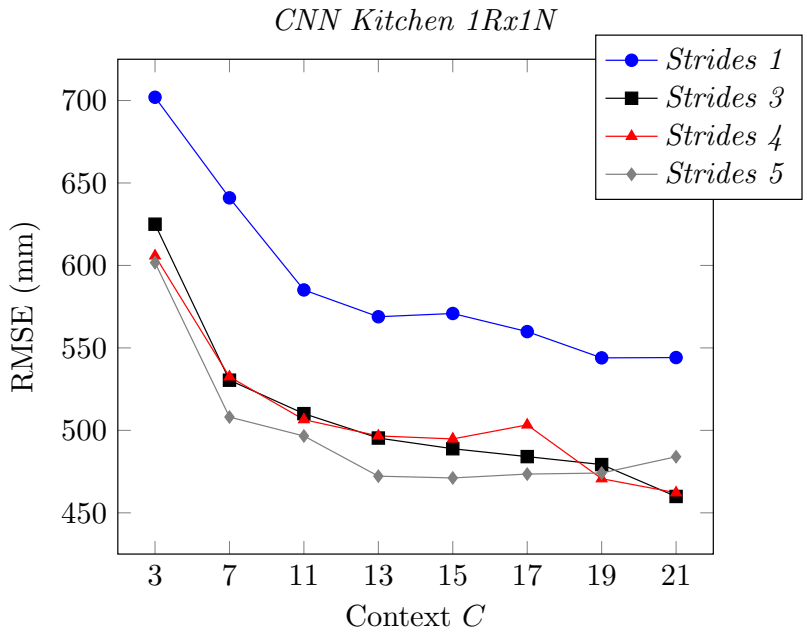
Ricerca : SLOC - Approccio I

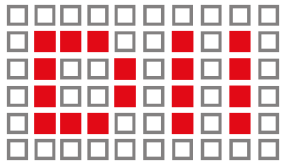
Advancements

- Paragonato con algoritmo Cross Spectrum Phase (CSP-SLOC) e Stereed Response Power (SRP-SLOC)

Algorithm	Average RMSE (mm)
CSP-SLOC	1464
SRP-SLOC	981
MLP-SLOC	406
CNN-SLOC	333

- Studio del contesto temporale



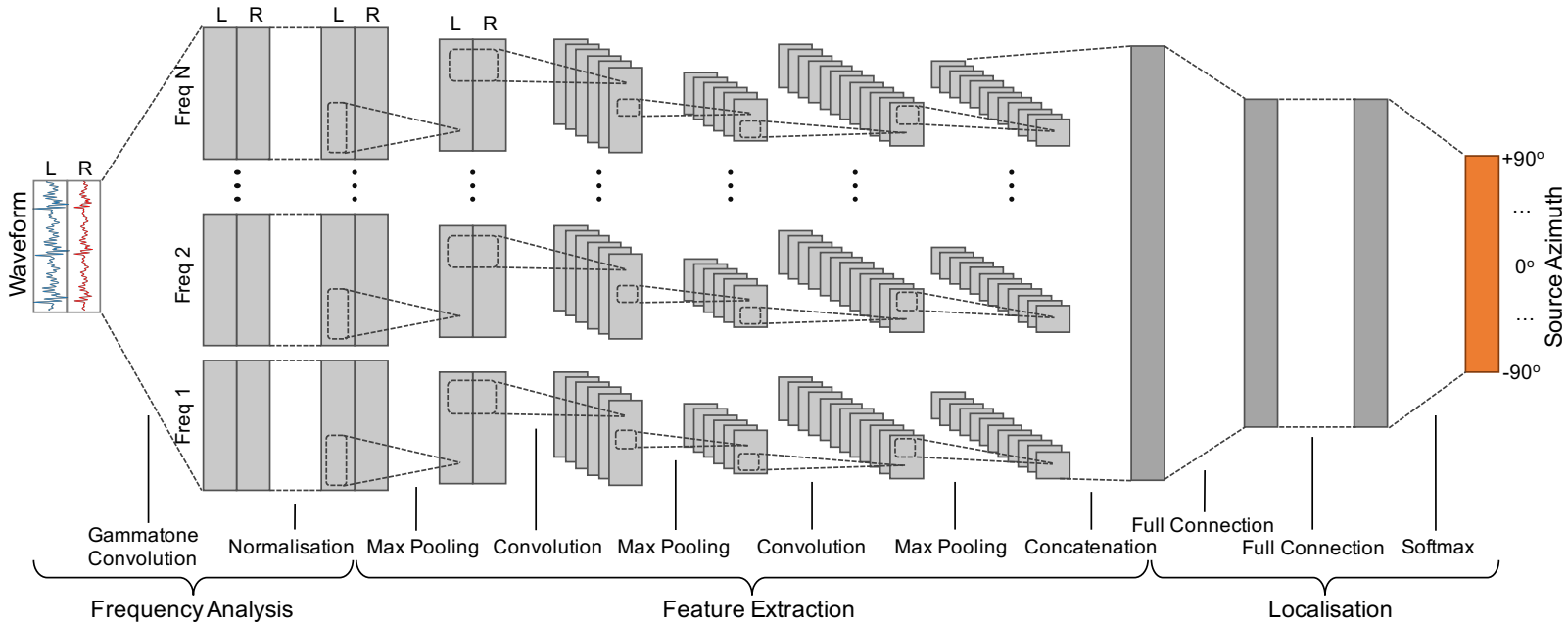


Elementi chiave:

- Determinare l'azimuth o l'altezza del parlatore
- Studio binaurale (manichino)
- Classificazione attuata da reti neurali
- Studio basato sull'apparato uditivo umano

Stima dell'azimuth del parlatore

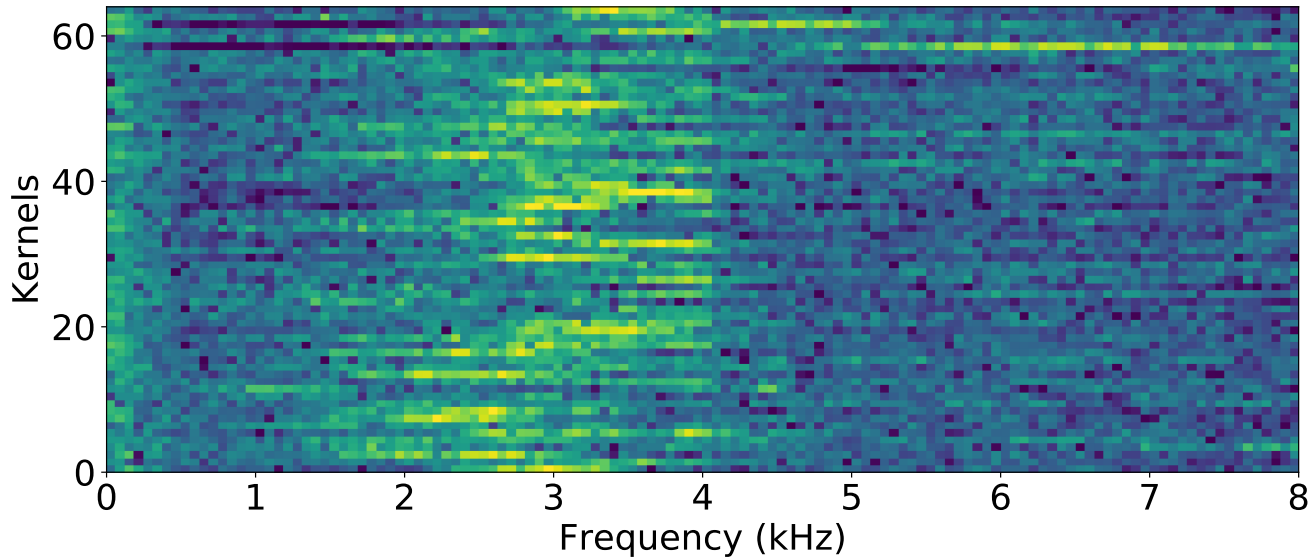
- Convolutional Neural Networks permettono filtraggio direttamente nel tempo
- Analogie con l'apparato uditivo umano - banco filtri Gammatone
- Approccio end-to-end
- Analisi in frequenza del segnale
- Primo modello in letteratura



Stima dell'azimuth del parlatore

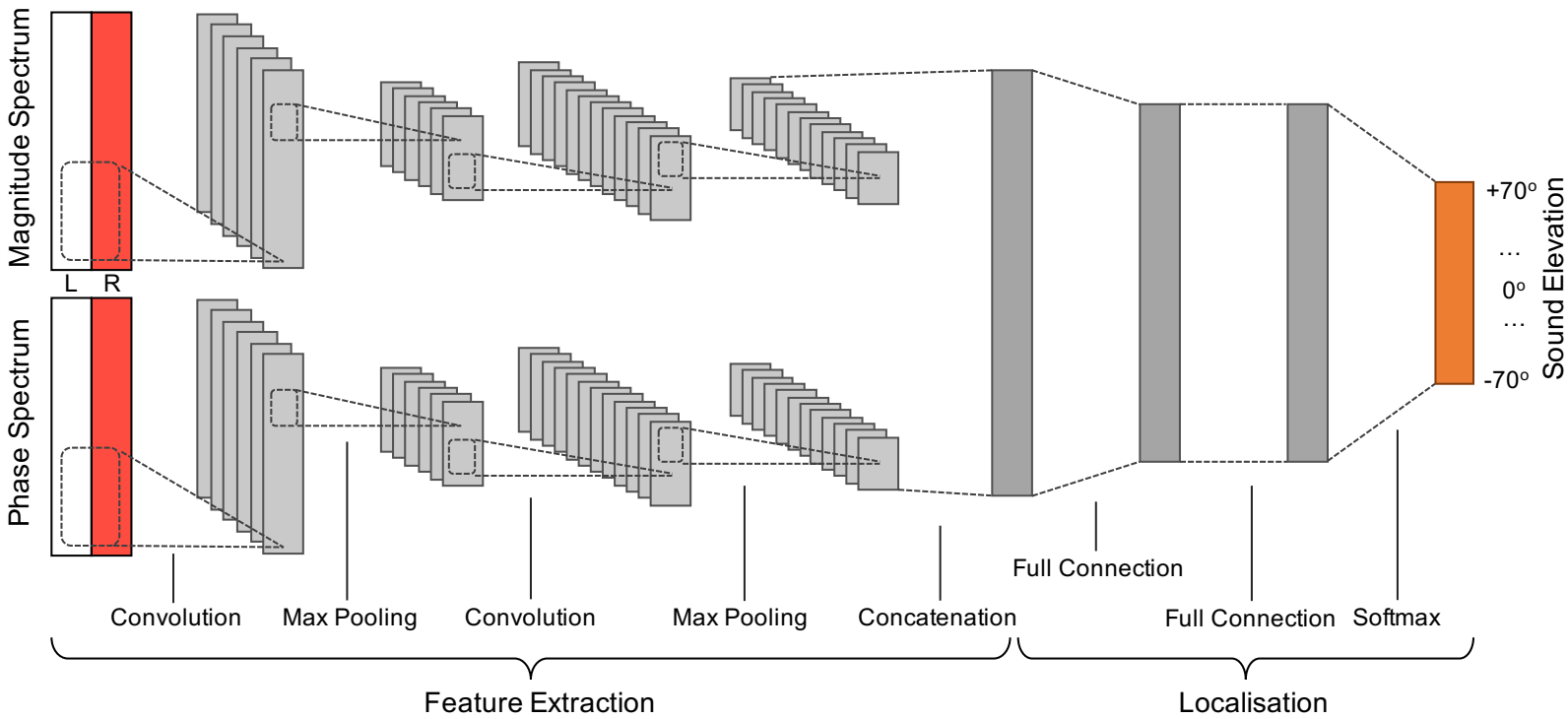
	Room A	Room B	Room C	Room D
Baseline	2.7°	3.3°	3.1°	5.2°
Wave-CONV	1.7°	2.3°	1.4°	2.4°
Wave-GTF	1.5°	3.0°	1.7°	3.5°

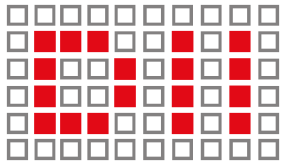
Studio del comportament in frequenza del modello *Wave-CONV*.



Stima dell'elevation del parlatore

- Convolutional Neural Networks
- Fase e ampiezza della FFT come input features





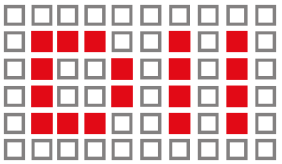
Stima dell'elevation del parlatore

Model	Anechoic		Room A		Room B		Room C	
	10 ms	100 ms	10 ms	100 ms	10 ms	100 ms	10 ms	100 ms
MFCC-CCF [1][§]	1.59deg		2.03deg		10.07deg		12.97deg	
GCC-PHAT	37.74deg	18.74deg	46.22deg	32.39deg	41.97deg	28.92deg	40.13deg	25.06deg
PHASE	11.18deg	0.83deg	18.38deg	7.09deg	14.12deg	3.16deg	12.37deg	2.03deg
MAG	6.20deg	0.34deg	11.12deg	2.33deg	8.40deg	1.28deg	7.24deg	0.89deg
PHASE-MAG	**3.40**deg	**0.00**deg	**6.78**deg	**0.18**deg	**4.47**deg	**0.03**deg	**3.75**deg	**0.05**deg

[§] O'Dwyer et al. [1] reported results with 1-s chunks.

1. H. O'Dwyer, E. Bates, and F. M. Boland,
"Machine learning for sound elevation detection,"
in *Proc. 4th Workshop on Intelligent Music Production*, Sep 2018.

Ricerca : Integrazione VAD e SLOC



Elementi chiave:

- Unico framework capace di fare VAD e SLOC simultaneamente
- Migliorare le performance del modello
- Utilizzo di più features
- Reti neurali convoluzionali

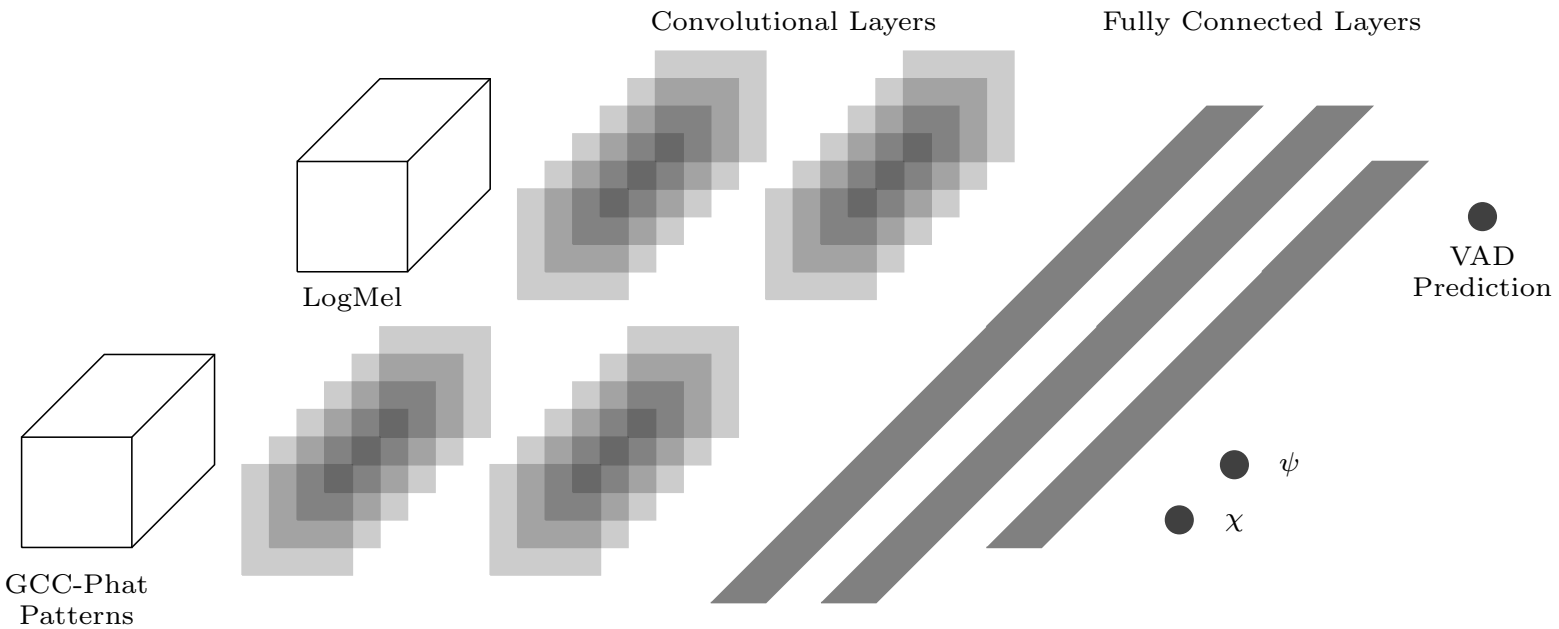
Problematiche:

- Ambiente multi-room
- Lo SLOC deve localizzare anche il silenzio
- Dipendenza dello SLOC dagli errori del VAD

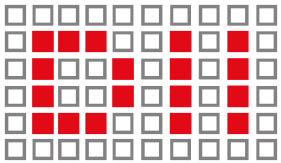
Ricerca : Integrazione VAD e SLOC

Studio Preliminare

- CNN con 2 input e 3 uscite
- Features: LogMel (detection) + GCC-PHAT (localization)
- Modello regressivo
- Non speech identificato come una posizione fuori dalla stanza
- Performance migliori in termini di VAD



Ricerca : Integrazione VAD e SLOC



Studio Preliminare

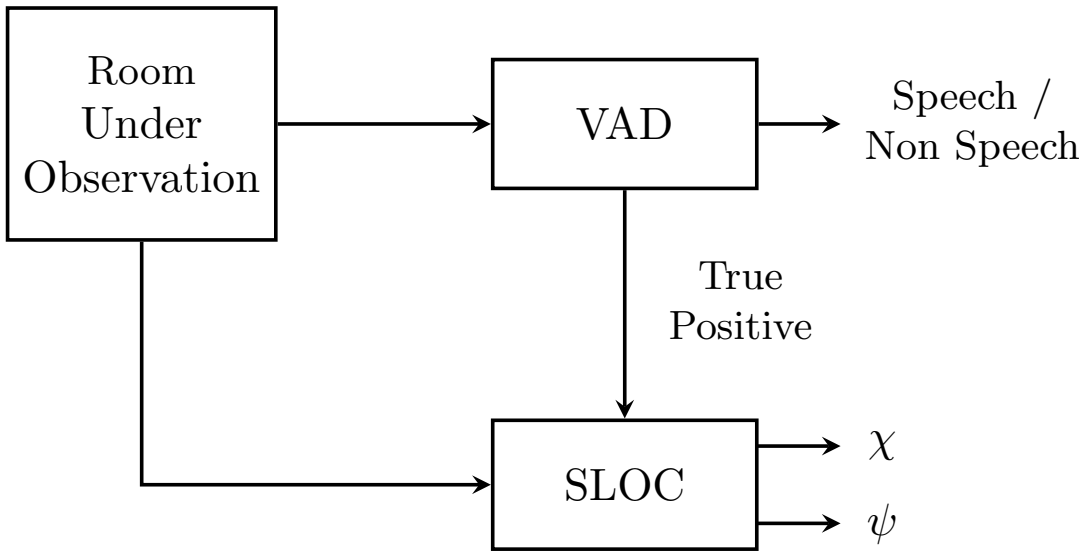
Detection	Neural VAD	Joint VAD-SLOC Model
SAD (%)	5.2	3.5
DEL (%)	6.2	4.2
FA (%)	4.2	2.8

Localization	Neural SLOC*	Joint VAD-SLOC Model	Neural SLOC†
RMS (mm)	327	629	318

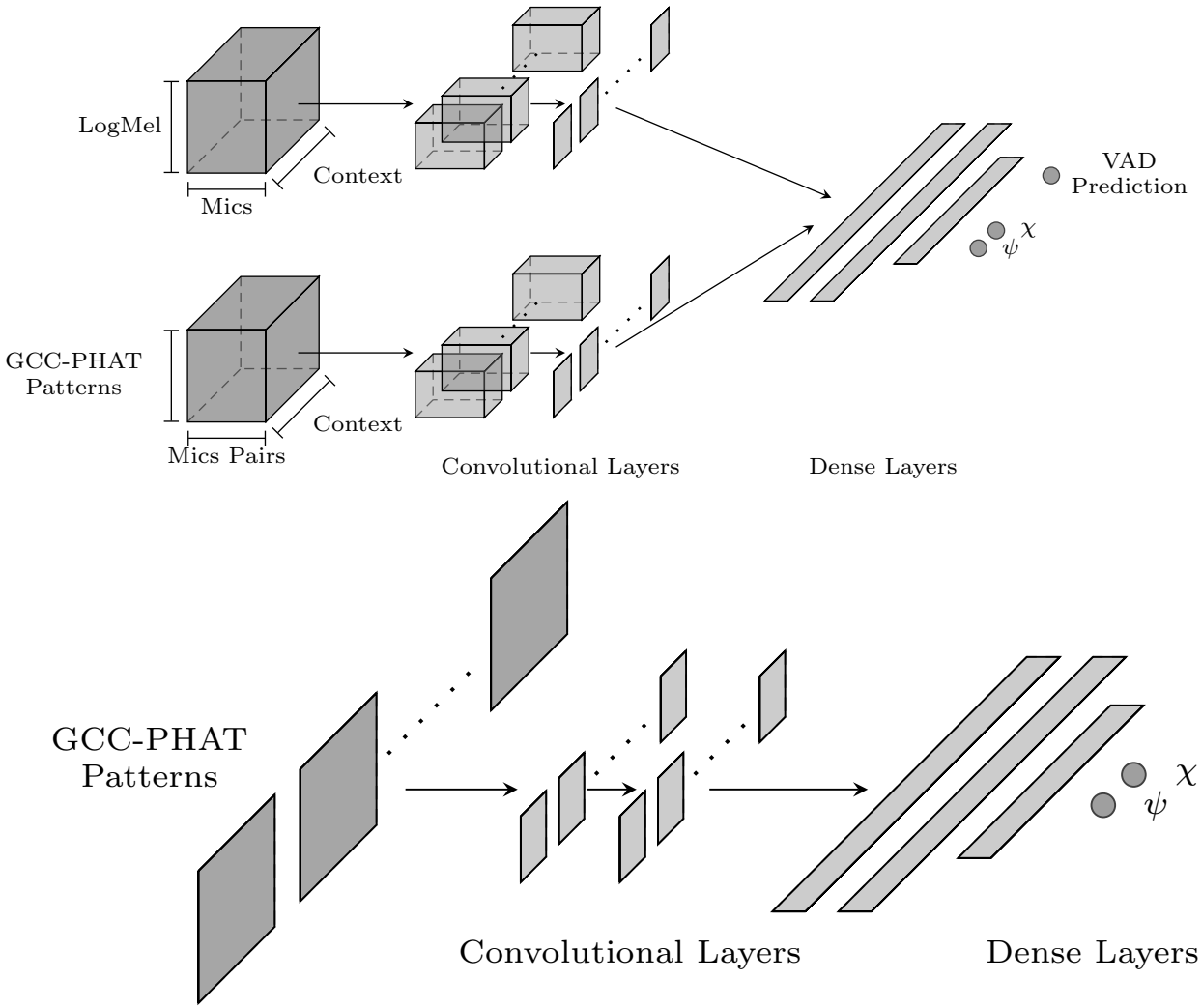
Ricerca : Integrazione VAD e SLOC

Advancements

- Nuovo modello proposto per localizzazione
- Confronto con unico modello presente in letteratura
- Data Augmentation
- SLOC in cascata al Joint VAD
- Localizzazione sui True Positive riconosciuti dal VAD



Ricerca : Integrazione VAD e SLOC



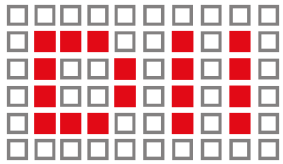
Ricerca : Integrazione VAD e SLOC

		Kitchen	Living Room	Average
Joint VAD	SAD (%)	10.1	12.4	11.0
	DEL (%)	16.4	20.4	18.4
	FA (%)	4.7	2.7	3.7
Joint VAD†	SAD (%)	6.3	5.3	5.8
	DEL (%)	11.3	9.1	10.2
	FA (%)	1.3	1.5	1.4

Oracle VAD		Average
Δ	RMS (mm)	-689

		Average
Δ	SAD (%)	-0.9
	DEL (%)	+4.1
	FA (%)	-4.7
	RMS (mm)	-640

Altri contributi: Filtraggio per crossover audio

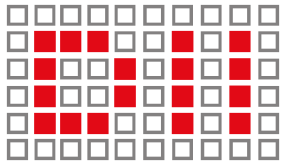


Design di Filtri IIR a fase quasi lineare tramite tecniche di intelligenza artificiale

Elementi chiave:

- Contesto automotive
- Crossover digitali per loudspeakers
- Crossover FIR: fase lineare, costo computazionale alto
- Crossover IIR: basso costo computazionale, fase non lineare, introdotti dei "buchi" in frequenza
- Sono stati proposti metodi per IIR a fase quasi lineare
- Utilizzo di tecniche di machine learning

Altri contributi: Filtraggio per crossover audio

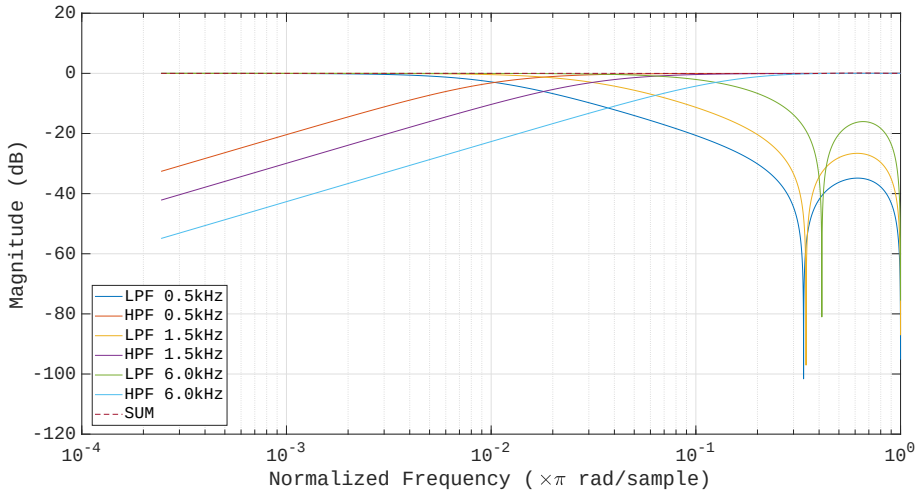


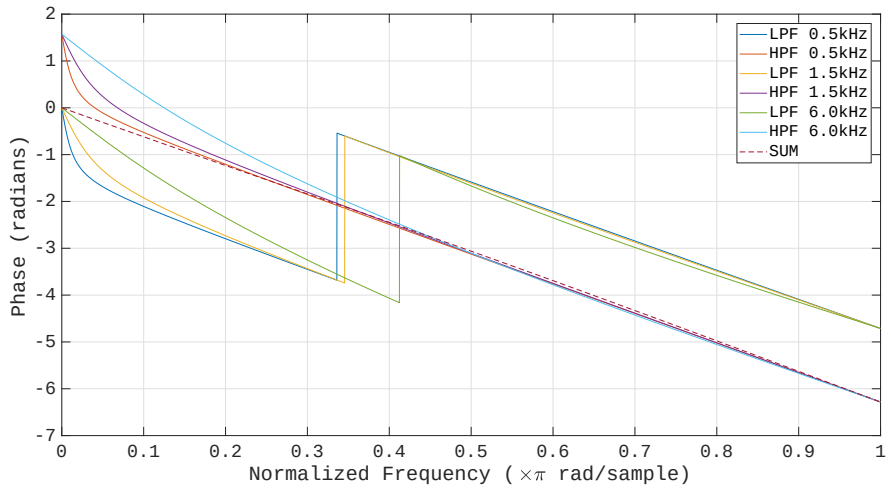
Metodo

- Uso della tecnica Fractional Derivative; permette di raggiungere la fase lineare
- Vincolo sulla frequenza di taglio del filtro
- Necessità di trovare i parametri ottimi della tecnica basata su FD
- Tecniche di intelligenza artificiale usate per trovare i parametri:
  - Artificial Bee Colony (ABC)
  - Particle Swarm Optimitazion (PSO)
- Confronto con tecnica senza vincolo sulla frequenza di taglio

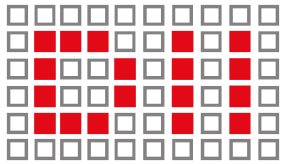
Altri contributi: Filtraggio per crossover audio

Risultati:





Attività di Formazione e Didattica



Lezioni Seguite:

- Economia e Management del Trasferimento Tecnologico
- Progettare la ricerca: i progetti europei

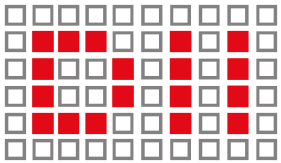
Attività Didattica di Supporto:

- Sviluppo delle slides per il corso Circuiti Algoritmi Elaborazione Segnali 2.

Attività di Tutor per i Tesisti:

- Ferdinando Foresi
- Giovanni Pepe

Partecipazione a conferenze



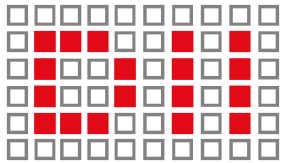
Conferenza internazionale:

- IEEE International Workshop on Machine Learning for Signal Processing, IEEE MLSP2016, Vietri sul Mare, 12-16 Settembre 2016. *[1 presentazione]*

- 144th Audio Engineering Society Pro Audio Convention, AES 2018, Milano, 23-26 Maggio 2018. *[1 poster]*

- 26th edition of the European Signal Processing Conference, EUSIPCO 2018, Roma, 03-07 Settembre 2018. *[1 poster]*

Lista Pubblicazioni



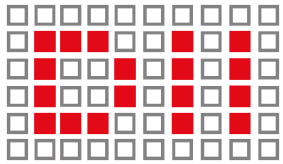
1. F. Vesperini, P. Vecchiotti, E. Principi, S. Squartini, and F. Piazza, "Deep Neural Networks for Multi-Room Voice Activity Detection: Advancements and Comparative Evaluation," *International Joint Conference on Neural Networks (IJCNN)*, pp. 3391-3398, 2016.

2. P. Vecchiotti, F. Vesperini, E. Principi, S. Squartini, and F. Piazza, "Convolutional Neural Networks with 3-D Kernels for Voice Activity Detection in a Multiroom Environment," *Italian Workshop on Neural Networks (WIRN*, 2016.

3. F. Vesperini, P. Vecchiotti, E. Principi, S. Squartini, and F. Piazza, "A Neural Network based algorithm for speaker localization in a multi-room environment," *Machine Learning for Signal Processing*, 2016.

4. F. Vesperini, P. Vecchiotti, E. Principi, S. Squartini, F Piazza, "Localizing Speakers in Multiple Rooms by Using Deep Neural Networks", Major Revision required in: *Computer Speech & Language, 2017*

Lista Pubblicazioni



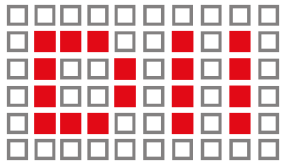
5. F. Foresi, P. Vecchiotti, D. Zallocco, and S. Squartini, "Designing Quasi-Linear Phase IIR Filters for Audio Crossover Systems by Using Swarm Intelligence," *Audio Engineering Society Convention 144*,2018.

6. P. Vecchiotti, E. Principi, S. Squartini, and F. Piazza, "Deep Neural Networks for joint Voice Activity Detection and Speaker Localization," *European Signal Processing Conference -26th Edition (EUSIPCO)*, 2018.

7. P. Vecchiotti, G. Pepe, E. Principi, and S. Squartini "A Deep Learning based method exploiting Data Augmentation for Joint Voice Activity Detection and Speaker Localization in residential environments," *Computer Speech and Language*, 2018, **Submitted**.

8. P. Vecchiotti, N. Ma, S. Squartini, and G. J. Brown "End-to-end Sound Localisation From the Raw Waveform," *International Conference on Acoustics, Speech, and Signal Processing*, 2019, **Submitted**.

Lista Pubblicazioni



9. N. Ma, P. Vecchiotti, and G. J. Brown "A Convolutional Neural Network for Estimating Sound Source Elevation in Reverberation Using Phase and Magnitude Spectra," *International Conference on Acoustics, Speech, and Signal Processing*, 2019, **Submitted**.