Reti CNN per l'elaborazione di segnali multicanale e localizzazione di sorgenti acustiche applicate a sistemi di sorveglianza intelligenti

Daniele Salvati, Carlo Drioli, Gian Luca Foresti
Dipartimento di Scienze matematiche, informatiche e fisiche, Università degli Studi di Udine
daniele.salvati@uniud.it, carlo.drioli@uniud.it, gianluca.foresti@uniud.it

Abstract

L'articolo tratta dell'applicazione di modelli di reti neurali convoluzionali (CNN) a problemi di localizzazione di sorgenti acustiche. Si discute il problema di stima della direzione di arrivo (DOA) in ambiente rumoroso e riverberante utilizzando un array lineare uniforme (ULA). Le reti CNN sono utilizzate per elaborare i dati multicanale dell'array e per migliorare lo schema di fusione dei dati eseguito nel calcolo delle risposte acustiche (steered response power, SRP). In particolare, le CNN permettono di pesare le diverse componenti di frequenza nella fusione dei dati in modo da ridurre gli effetti deleteri e gli artefatti dovuti al rumore e al riverbero. Esperimenti con dati reali dimostrano l'efficacia di questo approccio in termini di accuratezza della localizzazione. Questo metodo è applicato a sistemi di sorveglianza intelligenti, in cui la stima del DOA può essere usata per puntare una videocamera verso la sorgente o per rilevare e riconoscere eventi acustici di interesse.

1 Introduzione

Le tecniche di elaborazione di segnali multicanale hanno un ruolo centrale in applicazioni quali interazione uomo-computer, sistemi di teleconferenza, robotica, automazione, sorveglianza e svolgono processi importanti in una serie di applicazioni: localizzazione, separazione di sorgenti e miglioramento segnale-rumore, riconoscimento.

La localizzazione di sorgenti acustiche è stata ampiamente investigata dalla comunità scientifica e numerosi algoritmi sono stati sviluppati, come ad esempio le tecniche basate su time delay estimation o quelle di beamforming. Il beamforming è un filtro spaziale che opera sull'uscita di una schiera di microfoni allo scopo di enfatizzare o attenuare i segnali in funzione della direzione di provenienza fornendo risposte acustiche direzionali (steered response power, SRP), ed è una delle tecniche più robuste in condizioni di rumore e riverbero. Recentemente, è aumentato l'interesse nell'uso dei metodi di apprendimento automatico per la localizzazione acustica [Takeda e Komatani, 2016; Salvati *et al.*, 2016; Salvati *et al.*, 2018].

In questo articolo, si descrive l'impiego dei modelli convoluzionali nel contesto dell'elaborazione audio multicanale per la localizzazione acustica di una sorgente. Si considera la stima della direzione di provenienza del segnale (DOA) in ambito far-field, in condizioni rumorose e riverberanti, utilizzando un array lineare uniforme (ULA). Il sistema si basa su uno schema ibrido che integra le reti CNN nella catena di elaborazione incentrata sul beamforming minimum variance distortionless response (MVDR) [Capon, 1969]. L'algoritmo è basato su una rete CNN addestrata per classificare le componenti SRP a banda stretta in funzione della capacità di contribuire in modo costruttivo agli SRPs migliorando la localizzazione. Nella fase di fusione delle informazioni in frequenza le componenti che contribuiscono in modo negativo sono scartate o attenuate. Il modulo di localizzazione MVDR-CNN può essere integrato in un sistema di sorveglianza intelligente, e ha come obiettivo quello di rilevare sorgenti nel tempo e nello spazio fornendo informazioni alle videocamere PTZ che sono poi guidate verso la posizione di interesse per un'analisi audio-video della scena. Inoltre, la localizzazione è una informazione utile per elaborare ulteriormente il segnale sulla posizione stimata attraverso tecniche di separazione di sorgenti e miglioramento del segnale-rumore in modo da aumentare la capacità di riconoscimento del tipo di suono attraverso metodi avanzati di intelligenza artificiale.

2 Metodo

Il beamforming acustico è in genere compiuto nel dominio delle frequenze trasformando gli M segnali dei microfoni via short-time Fourier transform (STFT). Ogni componente a banda stretta è elaborata con il filtro MVDR fornendo la risposta direzionale in potenza per ogni DOA di interesse:

$$P(f,\theta) = \frac{1}{\mathbf{a}^H(f,\theta)\mathbf{\Phi}^{-1}(f)\mathbf{a}(f,\theta)}, \qquad (1)$$

dove f è l'indice di frequenza, $\mathbf{a}(f,\theta)$ è il vettore di puntamento dell'array che permette di sincronizzare e sommare i segnali verso la direziona angolare θ, e $\mathbf{\Phi}(f) = E[\mathbf{s}(f)\mathbf{s}^H(f)]$ è la matrice di covarianza dei segnali ($\mathbf{s}(f) = [S_1(f), S_2(f), ..., S_M(f)]$ è il segnale dell'array nel domino della frequenza).

Lo schema di fusione MVDR-CNN delle componenti a banda stretta si basa sulla seguente somma normalizzata

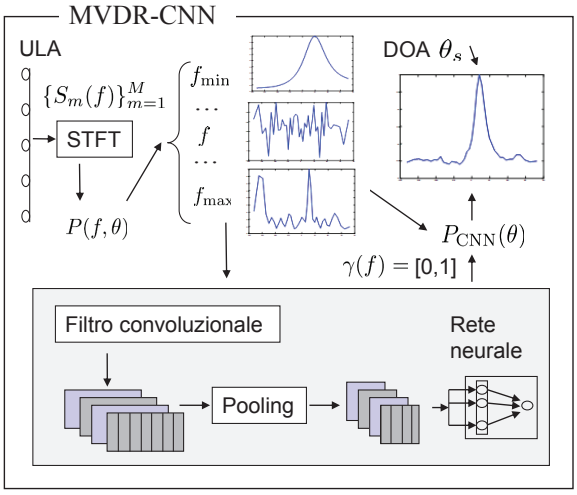

Figura 1: Architettura del sistema di elaborazione.

pesata:

$$P_{\text{CNN}}(\theta) = \sum_{f=f_{\min}}^{f_{\max}} \gamma(f) \frac{P(f,\theta)}{||\mathbf{p}(f)||_\infty}, \quad (2)$$

dove $f_{\min}$ e $f_{\max}$ sono i valori minimo e massimo dello spettro in frequenza considerato, $\mathbf{p}(f) = [P(f,\theta_1), P(f,\theta_2), \ldots, P(f,\theta_D)]$ è l'SRP per i D DOAs di interesse, e $\gamma(f)$ sono i fattori di peso calcolati dalla CNN, che possono assumere valori fra 0 e 1. La risposta direzionale in potenza che ha la massima energia corrisponde alla stima del DOA della sorgente: $\hat{\theta}_s = \underset{\theta}{\text{argmax}}[P_{\text{CNN}}(\theta)]$. La normalizzazione delle componenti in frequenza ha lo scopo di incrementare la risoluzione spaziale della fusione [Salvati *et al.*, 2014].

Si definisce quindi una funzione non lineare $F(\mathbf{p}(f), \mathbf{\Theta})$, basata su una rete CNN ($\mathbf{\Theta}$ sono i parametri acquisiti durante la fase di apprendimento), che mappa la SRP a banda stretta $\mathbf{p}(f)$ al peso $\gamma(f)$, il quale contribuisce ad enfatizzare o attenuare la componente nella fusione: $\gamma(f) = F(\mathbf{p}(f), \mathbf{\Theta})$. Il modello CNN è addestrato sulla base del seguente criterio di peso durante l'addestramento:

$$\overline{\gamma}_i(f) = \max\Big[1 - \frac{|\theta_s - \widehat{\theta}_s(f)|}{\eta}, 0\Big], \quad (3)$$

dove θ_s è il DOA delle sorgente, $\widehat{\theta}_s(f) = \underset{\theta}{\text{argmax}}[P(f,\theta)]$, e η è un valore di soglia per l'errore sull'angolo. L'architettura del sistema di elaborazione MVDR-CNN è riassunto in Figura 1. La struttura CNN ha un solo filtro convoluzionale di 20 kernels e dimensione 5×5 seguito da un modulo di max-pooling di dimensione 2×2.

3 Risultati

Il metodo MVDR-CNN è stato testato in due ambienti reali riverberanti utilizzando ULA con diverso numero di microfoni. Nel primo caso, un ULA di 3 microfoni è stato usato in una stanza di 6.37 m $\times$ 2.98 m $\times$ 3.6 m con tempo di riverberazione (RT_{60}) di 0.6 s. Un segnale vocale è stato riprodotto da un altoparlante in diverse posizioni nella stanza. Nel secondo caso, un ULA di 8 microfoni è stato usato in una stanza di 16 m $\times$ 7 m $\times$ 3 m con RT_{60} di 0.9 s. In questo test, sono stati registrati segnali da 3 differenti parlatori.

Tabella 1: AR (%) e RMSE (gradi) con un ULA di tre microfoni in una stanza di 6.37 m $\times$ 2.98 m $\times$ 3.6 m con RT_{60} di 0.6 s.

	MVDR	MVDR-CNN	SRP-PHAT
AR	91.13	**94.34**	85.96
RMSE	5.046	**4.625**	8.352

Tabella 2: AR (%) e RMSE (gradi) con un ULA di otto microfoni in una stanza di 16 m $\times$ 7 m $\times$ 3 m con RT_{60} of 0.9 s.

	MVDR	MVDR-CNN	SRP-PHAT
AR	71.29	**79.01**	67.05
RMSE	12.728	**6.898**	18.628

L'addestramento della rete CNN è stato effettuato in un ambiente rumoroso e riverberante simulato di dimensioni 5 m $\times$ 4 m $\times$ 3 m con RT_{60} di 0.5 s, usando un USASI noise come segnale della sorgente. La soglia η adottata è 5 gradi.

I risultati mostrati nelle Tabelle 1 e 2 indicano un miglioramento in termini di percentuali di accuratezza (AR, errore minore di 5 gradi) e di errore quadratico medio (RMSE) in comparazione al MVDR senza componete CNN, e al beamforming convenzionale con il filtro di normalizzazione phase transform (SRP-PHAT). L'MVDR-CNN risulta robusto alle variazioni dei tempi di riverberazioni e delle dimensioni della stanza rispetto la fase di addestramento.

Riferimenti bibliografici

[Capon, 1969] J. Capon. High resolution frequency-wavenumber spectrum analysis. *Proceedings of the IEEE*, 57(8):1408–1418, 1969.

[Salvati *et al.*, 2014] D. Salvati, C. Drioli, e G. L. Foresti. Incoherent frequency fusion for broadband steered response power algorithms in noisy environments. *IEEE Signal Processing Letters*, 21(5):581–585, 2014.

[Salvati *et al.*, 2016] D. Salvati, C. Drioli, e G. L. Foresti. A weighted MVDR beamformer based on SVM learning for sound source localization. *Pattern Recognition Letters*, 84:15–21, 2016.

[Salvati *et al.*, 2018] D. Salvati, C. Drioli, e G. L. Foresti. Exploiting CNNs for improving acoustic source localization in noisy and reverberant conditions. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2(2):103–116, 2018.

[Takeda e Komatani, 2016] R. Takeda e K. Komatani. Sound source localization based on deep neural networks with directional activate function exploiting phase information. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 405–409, 2016.