UN SISTEMA A COSTO MINIMO PER IL MIGLIORAMENTO QUALITATIVO DEL PARLATO

Danilo Comminiello[*], Aurelio Uncini[*], Albenzio Cirillo[†],
Antonino Barone[‡], Mauro Falcone[†]
[*] Dipartimento DIET, "Sapienza" Università di Roma - [†] Fondazione Ugo Bordoni - [‡] Iscom, MiSE
danilo.comminiello@uniroma1.it, aurel@ieee.org, albenzio.cirillo@gmail.com,
antonino.barone@sviluppoeconomico.gov.it, falcone@fub.it

1. ABSTRACT

Questo lavoro propone una tecnica innovativa di *beamforming* adattativo per applicazioni di miglioramento qualitativo del parlato nell'ambito di comunicazioni viva voce. In particolare, viene provata l'efficacia della tecnica per una schiera "minima", ovvero composta da soli due microfoni. Il contributo innovativo del sistema è apportato dall'algoritmo adattativo utilizzato, il *variable step-size block exact affine projection algorithm* (VSS-BEAPA), derivato dalla famiglia di algoritmi APA (*affine projection algorithms*), che si basa su un'elaborazione a blocchi e sull'utilizzo di un passo di adattamento variabile che permette di considerare scenari in cui la risposta impulsiva acustica viene sotto-modellata. La valutazione del sistema è condotta rispetto a diverse condizioni di lavoro e quindi rispetto a specifiche problematiche che possiamo così riassumere: presenza di rumore di fondo ambientale, presenza di sorgente interferente con il parlante, presenza di effetto eco dovuto al "ritorno di segnale" nella catena comunicativa. Le prestazioni sul miglioramento della qualità del segnale derivato dall'utilizzo del sistema proposto sono valutate attraverso la stima del rapporto segnale-rumore (SNR) e la stima oggettiva di un indice di intelligibilità. Gli esperimenti mostrano come il sistema proposto, pur non beneficiando di una schiera di sensori ottimale, riesca a produrre un sensibile decremento del livello di rumore interferente e un conseguente miglioramento qualitativo del segnale vocale.

2. INTRODUZIONE

L'apparato uditivo umano sfrutta la componente binaurale, ed in particolare i ritardi, ossia la fase dei segnali audio che arrivano alle nostre orecchie, per distinguere una specifica voce immersa in un contesto di rumori ed altre conversazioni sovrapposte a quella di nostro interesse. Questa caratteristica umana di saper separare i suoni nello spazio è stata studiata a fondo, cercando di riconoscere, e quindi ricreare, i processi che ne permettono il funzionamento, in modo da migliorare la "qualità" del segnale audio.

Effettuare una teleconferenza o, più in generale, far uso di sistemi viva voce senza alcun tipo di elaborazione del segnale acquisito dal microfono, è improponibile poiché la presenza di rumore o di sorgenti interferenti sovrapposte al segnale vocale principale renderebbe complicato preservare l'intelligibilità di quest'ultimo, compromettendo qualsiasi comunicazione.

Nel corso degli ultimi anni sono state sviluppate diverse tecniche che traggono origine dagli studi sull'apparato binaurale umano (Lotter & Vary, 2006; Luo & Uvacek, 2002; Stern et al., 2008). Utilizzando una schiera di microfoni è possibile elaborare i relativi segnali secondo le tecniche di *beamforming* (Benesty et al., 2010), ovvero combinandoli in

un'unica forma d'onda in cui risulta esaltato il segnale vocale proveniente da una specifica direzione e, pertanto, vengono attenuati i rumori provenienti dalle altre direzioni. Il beamforming può essere interpretato come il risultato di un filtro spaziale, in quanto prevede un trattamento del suono differente a seconda del punto nello spazio dove questo viene acquisito. Gli studi presenti in letteratura (Benesty et al., 2010) mostrano come l'efficacia del beamforming cresca all'aumentare del numero di microfoni utilizzati nella schiera microfonica. Tuttavia, vi è ancora necessità di creare algoritmi di beamforming più robusti alle variazioni dovute allo spostamento del parlatore e alle condizioni di rumore.

Un classico sistema di beamforming è il *Generalized Sidelobe Canceller* (GSC) (Griffiths & Jim, 1982), composto da un beamformer fisso di tipo *delay-and-sum* (DSB), che ha lo scopo di focalizzare la sorgente vocale principale, e da un blocco di *cancellazione adattativa del rumore* (*adaptive noise canceller*, ANC), che riduce la potenza del rumore di fondo nel segnale in uscita al beamformer.

Il blocco ANC nelle applicazioni viva voce comporta l'utilizzo di filtri digitali FIR dell'ordine delle centinaia o anche migliaia, ed i cui coefficienti devono essere stimati adattativamente in modo da ridurre il rumore quanto più possibile. La scelta dell'algoritmo adattativo è la parte critica di un sistema GSC, in quanto deve essere garantita una stima rapida ed efficace dei valori del filtro. Generalmente, l'adattamento nel dominio del tempo nell'ANC viene effettuato tramite algoritmi basati sulla minimizzazione dell'errore quadratico medio, come il *least mean square* (LMS) e il *normalized LMS* (NLMS). Tuttavia questi algoritmi mostrano una lenta convergenza per filtri di elevata dimensione (Sayed, 2008; Uncini, 2010), a tal punto che l'adattamento diventa impraticabile in tempo reale. La famiglia di algoritmi basati sulla proiezione affine (*affine projection algorithms*, APA) (Ozeki & Umeda, 1984) mostra invece velocità di convergenza più alta, e complessità computazionale gestibile, motivo per cui l'APA è stato spesso utilizzato in applicazioni di beamforming adattativo (Zheng & Goubran, 2000).

L'algoritmo adattativo proposto nel nostro sistema è il *variable step size block exact affine projection algorithm* (VSS-BEAPA) (Comminiello et al., 2010). Il VSS-BEAPA è l'esatta trasposizione nel dominio della frequenza dell'algoritmo APA, potenziato da un passo di adattamento variabile che permette di considerare scenari dove la risposta impulsiva viene sotto-modellata, ossia la lunghezza del filtro risulta più piccola della reale lunghezza della risposta impulsiva, cosa che spesso accade nelle applicazioni viva voce.

Il sistema proposto è dunque formato da un'interfaccia microfonica "a costo minimo" ed un beamformer GSC con filtro VSS-BEAPA. Inoltre sono valutati anche i benefici derivanti dall'inserimento di un *post-filter* che ha lo scopo di ridurre ulteriormente il rumore di fondo presente nel segnale acquisito.

La valutazione del sistema è condotta rispetto a diverse condizioni di lavoro e, quindi, rispetto a problematiche specifiche, tali per cui il sistema è stato sottoposto alla presenza di rumori di varia natura, dal semplice rumore di fondo ambientale a rumori colorati di tipo *automotive* e *cocktail party*.

Il lavoro è organizzato nel seguente modo: nel Paragrafo 3 viene presentato il sistema di beamforming utilizzato, mentre l'algoritmo VSS-BEAPA viene descritto nel Paragrafo 4. Il Paragrafo 5 contiene un'ampia descrizione degli scenari sperimentali, delle analisi prestazionali e dei risultati ottenuti. Infine, nel Paragrafo 6 sono riportate le conclusioni finali.

3. SISTEMA DI RIDUZIONE DEL RUMORE

Il sistema di beamforming adattativo utilizzato è mostrato in Fig. 1, in cui è possibile notare come il sistema sia composto da un'interfaccia microfonica, un percorso fisso di DSB, e un percorso adattativo di cancellazione dei lobi laterali, in una tipica configurazione GSC. Infine, sull'uscita ottenuta viene applicato un post-filter con l'obiettivo di un'ulteriore riduzione del rumore. L'assunzione principale che viene fatta riguardo la *direzione di arrivo* (*direction of arrival*, DOA) del segnale desiderato, è che la sorgente desiderata si trovi nel "campo vicino" della schiera microfonica, mentre le sorgenti rumorose si trovino indistintamente in "campo lontano". Tuttavia l'assunzione fatta non comporta costrizioni particolari poiché descrive uno scenario tipico delle comunicazioni viva voce. Analizziamo ora in dettaglio le varie componenti del sistema utilizzato.

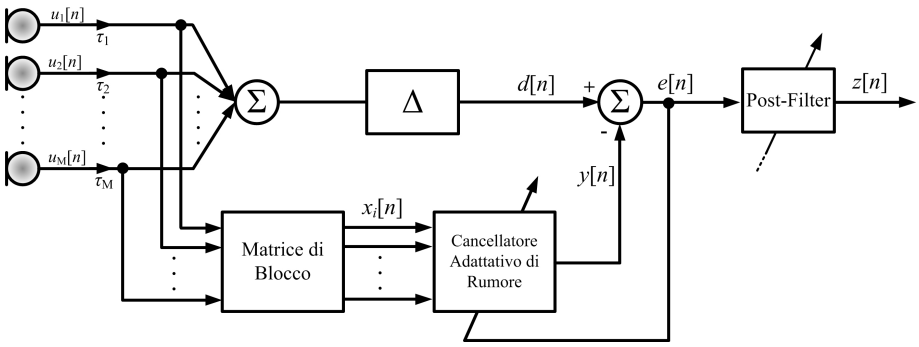

Figura 1: Sistema di beamforming adattativo.

3.1. L'interfaccia microfonica

La prima importante componente di un sistema di beamforming è costituita dall'interfaccia microfonica. Si consideri una schiera microfonica composta da un numero M di sensori. Il segnale $u_m[n]$ acquisito dall'm-esimo microfono, con $m = 1, \ldots, M$, contiene una replica ritardata del segnale obiettivo $s[n]$, ossia quello proveniente dalla sorgente desiderata, convoluto con la *risposta impulsiva acustica* $\mathbf{a}_m$ relativa al percorso che va dalla sorgente desiderata al corrispondente m-esimo microfono, con l'aggiunta di eventuale rumore ambientale di fondo $v_m[n]$.

La scelta della geometria dell'interfaccia microfonica gioca un ruolo fondamentale nel recupero della percezione binaurale. Una schiera microfonica ottimale per applicazioni di *speech enhancement*, ossia di miglioramento qualitativo del parlato, dovrebbe avere un'ampia apertura in modo da ottenere una buona risoluzione spaziale e allo stesso tempo evitare i fenomeni di *aliasing* che possono generare ambiguità nella distinzione e nell'acquisizione del segnale desiderato (Benesty, 2010). Questo compromesso viene raggiunto scegliendo in modo opportuno il numero di microfoni e la distanza fra essi.

In questo lavoro verrà proposta una tecnica originale di beamforming adattata ad una schiera "minima" ovvero di soli due microfoni, in modo da valutare il sistema anche in condizioni sfavorevoli. Ovviamente utilizzando un numero maggiore di sensori si otterrebbero risultati migliori, tuttavia la nostra scelta è anche dovuta alla facilità con cui oggigiorno è possibile usufruire di sistemi viva voce commerciali con due soli microfoni.

3.2. Il cancellatore di lobi laterali

Il GSC è il cuore del sistema. In esso, il DSB allinea spazialmente i segnali microfonici in base alla direzione della sorgente desiderata, generando un riferimento vocale desiderato $d[n]$. Il ramo adattativo riceve in ingresso i segnali microfonici $u_m[n]$ generando i riferimenti rumorosi $x_i[n]$, con $i = 1, \ldots, M\text{-}1$, per mezzo di una matrice di blocco che lascia passare tutte le componenti al di fuori della direzione della sorgente desiderata. Questi segnali vengono quindi filtrati dal cancellatore adattativo di rumore che rimuove la correlazione fra la componente di rumore presente nel riferimento vocale e i riferimenti rumorosi generati nel ramo adattativo, generando così il segnale di uscita $e[n]$ del beamformer contenente una sensibile riduzione del rumore.

Questa struttura sfrutta la configurazione della schiera microfonica massimizzando l'indice di direttività nella direzione desiderata e riducendo i segnali interferenti derivati dal rumore in campo diffuso.

3.3. Post-filter adattativo

I sistemi di cancellazione adattativa del rumore possono presentare una forte diminuzione delle prestazioni quando la sorgente di rumore non può essere modellata perfettamente come un singolo punto-sorgente (Bitzer et al., 1999). Questo è il motivo per cui si incontrano enormi difficoltà quando la sorgente è non-stazionaria oppure quando l'ambiente è riverberante. In questi casi infatti, i segnali di rumore interferenti giungono alla schiera microfonica da direzioni diverse a causa delle riflessioni all'interno dell'ambiente.

Allo scopo di migliorare le prestazioni del sistema, è stato aggiunto al beamformer un post-filter adattativo. Questo filtro aggiuntivo si basa sulle misurazioni a corto raggio delle funzioni di autocorrelazione e cross-correlazione dei segnali microfonici (Marro et al., 1998). In paricolare è stato applicato un post-filter adattativo progettato appositamente per il miglioramento qualitativo del segnale vocale in scenari di rumore in campo diffuso (McCowan & Bourlard, 2002). L'intero sistema trae beneficio da alcune importanti caratteristiche del post-filter. Per prima cosa, si ottiene una completa cancellazione delle componenti incoerenti del segnale, così che ci si può attendere un'elevata riduzione del riverbero. Inoltre, data la natura tempo variante del post-filter, è possibile ottenere una buona riduzione del rumore anche in ambienti acustici non stazionari.

4. IL FILTRO ADATTATIVO LINEARE VSS-BEAPA

4.1. Descrizione dell'algoritmo VSS-BEAPA

Il blocco di cancellazione adattativa del rumore ricopre un ruolo fondamentale nel processo di beamforming. L'ANC può essere visto come un sistema MISO (*multiple-input single-output*) composto da un banco di filtri adattativi, ciascuno relativo a un segnale di riferimento del rumore $x_i[n]$. La cancellazione dell'uscita di ciascun filtro dal segnale di riferimento vocale $d[n]$, generato dal DSB, produce un segnale di stima del rumore $y[n]$, che rappresenta la somma dei contributi sottratti al riferimento vocale, e un segnale di uscita del beamformer $e[n]$, che rappresenta il segnale di errore del processo adattativo.

Nelle applicazioni di *speech enhancement* l'ordine del filtro adattativo può essere molto grande richiedendo così un elevato costo computazionale. Generalmente, un buon compromesso fra complessità e prestazioni si ottiene utilizzando la famiglia di *algoritmi della proiezione affine* (APA), che nel beamforming adattativo consentono di ottenere velocità di convergenza più alte con una moderata complessità computazionale.

Il *variable step size block exact affine projection algorithm* (VSS-BEAPA) (Comminiello et al., 2010) deriva dall'implementazione in frequenza dell'APA a blocchi (Tanaka et al., 1999) con l'aggiunta di un passo di adattamento variabile. Ogni iterazione dell'algoritmo fornisce un blocco di P campioni del segnale di uscita del beamformer $e[n]$. Indichiamo con $b = 1, \ldots, B$, l'indice di blocco, dove B è il numero di blocchi. L'uscita del beamformer relativa al blocco b è un vettore $P\times1$ definito come:

$$\mathbf{e}_n^{[b]} = \mathbf{d}_n^{[b]} - \mathbf{y}_n^{[b]} \tag{1}$$

dove il vettore blocco $\mathbf{d}_n^{[b]}$ è una selezione del vettore di uscita del DSB avente lunghezza L ed è definito come: $\mathbf{d}_n = [d[n], d[n-1], ..., d[n-L+1]]^T$. Analogamente, $\mathbf{y}_n^{[b]}$ è un blocco del segnale di uscita dell'ANC, ottenuto come:

$$\mathbf{y}_n^{[b]} = \sum_{i=1}^{M-1} \mathbf{X}_{i,n}^T \mathbf{w}_n \tag{2}$$

Nell'equazione (2), M è il numero di microfoni e $\mathbf{X}_{i,n}$ è la matrice dei dati, di dimensione $L\times P$, contenente i riferimenti di rumore; $\mathbf{X}_{i,n}$ è generata utilizzando un ordine di proiezione P per ciascun segnale di riferimento del rumore ed è definita come:

$$\mathbf{X}_{i,n} = \begin{bmatrix} x_i[n] & x_i[n-1] & \text{L} & x_i[n-P+1] \\ x_i[n-1] & x_i[n-2] & \text{L} & x_i[n-P] \\ \text{M} & \text{M} & \text{O} & \text{M} \\ x_i[n-L+1] & x_i[n-L] & \text{L} & x_i[n-P-L+2] \end{bmatrix} \tag{3}$$

In (2), il vettore $\mathbf{w}_n$ di dimensione $L\times1$ contiene i coefficienti del filtro adattativo. Per ciascun microfono, l'equazione di aggiornamento risultante dell'algoritmo VSS-BEAPA è:

$$\mathbf{w}_n = \mathbf{w}_{n-1} + \mu[n]\mathbf{X}_{i,n}\mathbf{R}_n^{-1}\mathbf{e}_n^{[b]} \tag{4}$$

in cui $\mathbf{R}_n = \mathbf{X}_{i,n}^T\mathbf{X}_{i,n} + \delta\mathbf{I}$ è la matrice di covarianza dell'ingresso, di dimensioni $P\times P$, calcolata in modo ricorsivo (Sayed, 2008; Uncini, 2010); δ è un parametro di regolarizzazione, mentre $\mathbf{I}$ è una matrice identità $P\times P$. Nelle equazioni (2) e (4) le convoluzioni vengono implementate nel dominio della frequenza tramite trasformate veloci di Fourier (FFT), come in (Tanaka et al., 1999). In (4), $\mu[n]$ è il passo di adattamento tempo variante. L'utilizzo di un passo di adattamento variabile ci consente di scegliere una lunghezza del filtro L_F minore della lunghezza necessaria per il modellamento della risposta impulsiva acustica L, senza che venga introdotta alcuna componente irriducibile di rumore all'uscita del sistema. Il passo di adattamento variabile utilizzato deriva da un processo di minimizzazione della deviazione media quadratica (Paleologu et al., 2008):

$$\mu[n] = \begin{cases} \mu_\text{f}\,, & n \le L_\text{F} \\ \left| 1 - \dfrac{\sqrt{\left|\hat{\sigma}_s^2[n] - \hat{\sigma}_y^2[n]\right|}}{\hat{\sigma}_e^2[n] + \xi} \right|, & n > L_\text{F} \end{cases} \tag{5}$$

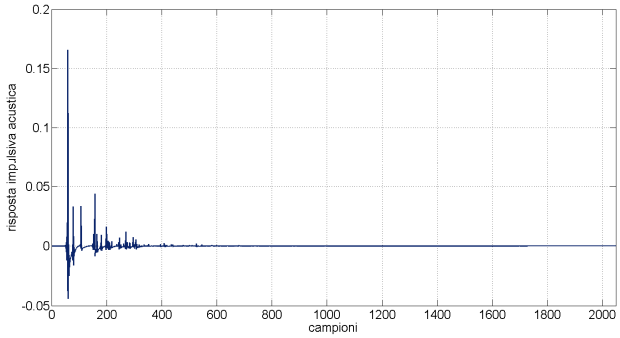

Figura 2: Risposta impulsiva acustica.

dove ξ è una piccola costante positiva che serve ad evitare divisioni per zero. Il parametro generico $\hat{\sigma}_{\alpha}^{2}[n]$, dove $\alpha = \{s, y, e\}$, rappresenta la stima di potenza della corrispondente sequenza generica $\alpha[n]$, e può essere calcolato come:

$$\hat{\sigma}_{\alpha}^{2}[n] = \lambda \hat{\sigma}_{\alpha}^{2}[n-1] + (1-\lambda)\alpha^{2}[n] \tag{6}$$

in cui λ è un fattore di *smoothing* scelto in modo che $\lambda = 1 - 1/(K N_F)$, con $K > 1$. Dato che per le prime L_F iterazioni il filtro non è sotto-modellato, il passo di adattamento è inizialmente fissato ad un valore μ_f e diventa variabile quando l'istante temporale $n > L_F$.

Osservando le equazioni (4) e (5) è possibile notare che l'algoritmo VSS-BEAPA utilizza esclusivamente parametri disponibili all'uscita del filtro adattativo, i.e. $s[n]$, $y[n]$, ed $e[n]$. Ciò implica che tutte le informazioni riguardanti la non stazionarietà acustica sono contenute nell'equazione (5). Questa caratteristica rende l'algoritmo VSS-BEAPA robusto a livelli di rumore ambientale anche molto elevati.

Rispetto all'algoritmo convenzionale NLMS o all'algoritmo APA a blocchi calcolato nel dominio del tempo, il VSS-BEAPA mostra prestazioni di convergenza migliori. Il VSS-BEAPA ottiene inoltre una considerabile riduzione dei costi computazionali oltre ad una latenza ridotta, dovuta principalmente all'elaborazione in blocchi nel dominio della frequenza e alla possibilità di sotto-modellare la risposta impulsiva acustica.

4.2. Valutazione prestazionale del filtro VSS-BEAPA

Allo scopo di valutare l'efficacia dell'algoritmo VSS-BEAPA, è stata effettuata una serie di esperimenti in ambiente simulato per misurare la bontà del filtro in condizioni ambientali rumorose. In particolare, è stato simulato uno scenario di teleconferenza in una tipica stanza di ufficio, in cui l'ambiente acustico può subire delle variazioni dovute alla non stazionarietà delle sorgenti presenti oppure ad una alterazione delle condizioni ambientali. Il tempo di riverbero dell'ambiente simulato è di circa T_{60} = 250 ms. Il segnale obiettivo è un rumore bianco a cui viene aggiunto, come rumore di fondo, un ulteriore rumore Gaussiano a media nulla e varianza unitaria, tale da produrre un *rapporto segnale-rumore* (SNR) in ingresso pari a 20 dB. La risposta impulsiva acustica, rappresentata in Fig. 2, viene calcolata in ambiente simulato tramite Roomsim (Campbell et al., 2005), uno strumento di MATLAB®, utilizzando una frequenza di campionamento di 8 kHz.

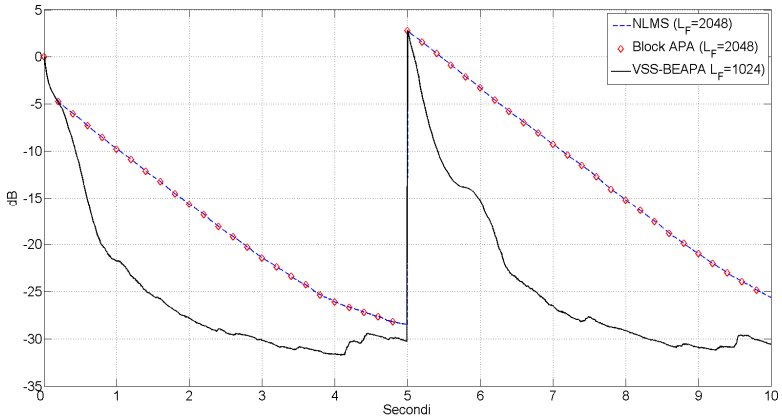

Figura 3: Andamento del disallineamento normalizzato.

La lunghezza della risposta impulsiva acustica è di L = 2048 campioni. Per quanto riguarda i valori dei parametri dell'algoritmo VSS-BEAPA, la nostra scelta è la seguente: K = 2, $\xi = 10^{-4}$, $\delta = 10^{-5}$, μ_f = 0.2. Scegliamo un ordine di proiezione pari a P = 2 e una lunghezza del filtro di L_F = 1024 campioni tale cioè di avere un sotto-modellamento pari alla metà della reale lunghezza delle risposta impulsiva acustica. La lunghezza di un esperimento ha una durata di 10 secondi. Allo scopo di simulare un cambio repentino nell'ambiente acustico, ossia una non stazionarietà, trasliamo verso destra la risposta impulsiva acustica di 20 campioni dopo 5 secondi dalla durata dell'esperimento.

Confrontiamo l'algoritmo utilizzato con l'algoritmo convenzionale NLMS e con l'algoritmo APA a blocchi. Per avere una valutazione qualitativa delle prestazioni dell'algoritmo utilizziamo come misura il *disallineamento normalizzato* $\mathcal{M}_s$, espresso in dB, definito per scenari sotto-modellati come:

$$\mathcal{M}_s = 20\log_{10}\left(\frac{\left\|\mathbf{h}_n - \hat{\mathbf{h}}_n\right\|_2}{\left\|\mathbf{h}_n\right\|_2}\right) \qquad (7)$$

dove $\mathbf{h}_n$ è il vettore della risposta impulsiva acustica e $\hat{\mathbf{h}}_n$ è il filtro stimato con aggiunta di zeri nella parte sotto-modellata. L'andamento del confronto fra i disallineamenti normalizzati dei vari algoritmi è rappresentato in Fig. 3, in cui è possibile notare che mentre l'NLMS e l'APA a blocchi nel dominio del tempo mostrano approssimativamente lo stesso andamento, il VSS-BEAPA ha un disallineamento minore sia in fase di convergenza sia a regime. Inoltre, il VSS-BEAPA reagisce anche più velocemente degli altri algoritmi ai cambiamenti della risposta impulsiva acustica.

5. RISULTATI SPERIMENTALI

In questa sezione andremo ad effettuare una valutazione del sistema di beamforming presentato utilizzando una schiera microfonica minima composta da due soli sensori.

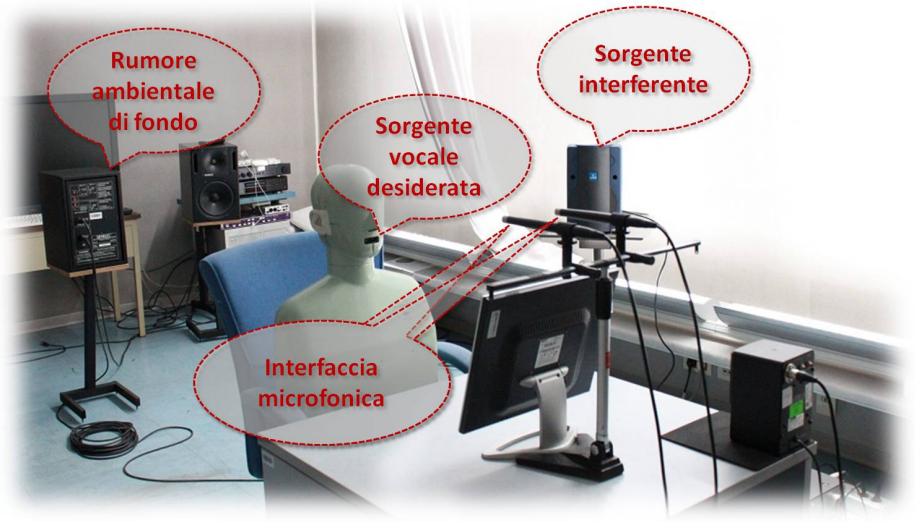

Figura 4: Set-up sperimentale.

5.1. Il set-up sperimentale

Lo scenario sperimentale è quello di una stanza silente allestita con diversi materiali, come tende, tavoli, ecc, al fine di ricreare un tipico ambiente di lavoro. Il set-up sperimentale è illustrato in Fig. 4 ed è composto da una sorgente vocale desiderata posta nel campo vicino dell'interfaccia microfonica e generata per mezzo di un busto artificiale, una sorgente interferente posta lateralmente e rivolta anch'essa verso la schiera microfonica, e infine una sorgente in campo diffuso che riprodurrà rumore ambientale di fondo di varia natura, allo scopo di ricreare un'ampia varietà di condizioni di lavoro. Sia la sorgente interferente sia la sorgente di rumore ambientale sono costituite da sistemi professionali di diffusione acustica.

Per quanto riguarda la parte di riproduzione dei segnali si è utilizzato: un diffusore amplificato GENELEC 1031 A per la generazione del rumore di fondo; un diffusore amplificato LEM Sound Pressure per la generazione del segnale interferente; una bocca artificiale 4227 inserita nel torso Bruel and Kjaer 4128. Per la parte di acquisizione invece il set microfonico Bruel & Kjaer 3529 con preamplificatori Bruel & Kjaer 2812 è collegato ad una Focusrite Voicemaster che opera un blando filtro alle basse frequenze (~100Hz) e amplifica i singoli canali al fine di un corretto allineamento. La conversione A/D dei segnali in uscita e la conversione D/A del segnale stereo è operata con la scheda Focusrite Saffire 40 PRO comandata dal software Adobe Audition 3.0. Al fine di mantenere le condizioni di registrazioni stabili, tutte le operazioni sono state eseguite tramite controllo remoto di tutte le strumentazioni di riproduzione e di registrazione del segnale.

I segnali vocali desiderati riprodotti dal busto artificiale fanno parte del database CLIPS, rilasciato nel 2004 e relativo ad un progetto finanziato dal MIUR. In particolare, è stata utilizzata la parte del database relativa al segnale ortofonico realizzata da parlatori professionisti in camera anecoica. Nel nostro caso, per ciascun parlatore è stato realizzato un segnale costituito da una sequenza di quattro frasi, ciascuna di 4 secondi circa.

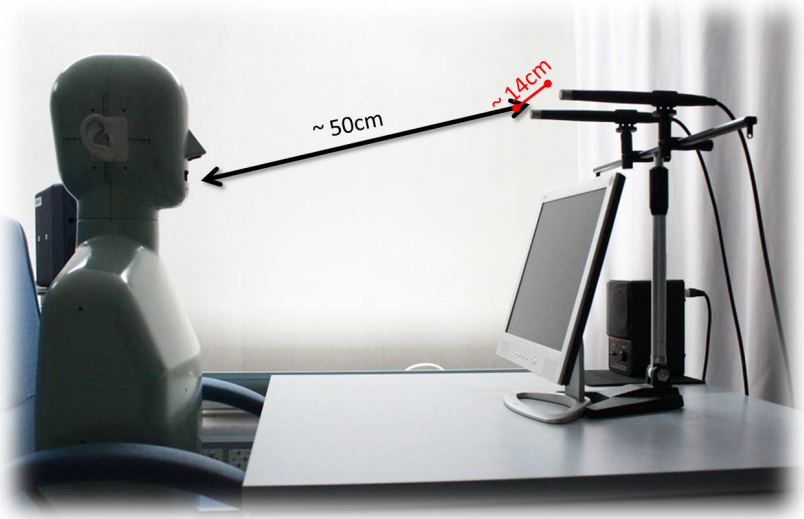

Figura 5: Posizionamento della sorgente desiderata e della schiera "minima".

Il rumore di fondo è stato realizzato invece utilizzando i segnali del CD NOISE-ROM-0 prodotto nell'ambito del progetto Europeo SAM nel 1990; sono stati utilizzati solo alcuni tra i più comuni rumori di fondo ovvero: rumore rosa, di autoveicolo, di fabbrica, e di effetto "*cocktail party*", o anche detto "*babble noise*", tipicamente generato da due o più persone che parlano tra loro generando un rumore di fondo molto fastidioso per l'intelligibilità di una comunicazione viva voce. Come segnale interferente invece, è stato utilizzato un segnale vocale tratto da un notiziario televisivo della emittente LA7. A tutti i segnali è stato anteposto un tono di calibrazione di 1 kHz per verificare il corretto livello dei segnali anche nelle registrazioni dopo l'allineamento e la calibrazione dell'intera catena di registrazione.

Un fattore di rilevante importanza per le prestazioni del sistema è ovviamente quello del posizionamento microfonico. Ci si è posti in quelle che potremo definire normali condizioni di lavoro con un posizionamento dei microfoni a circa 50 cm dal parlatore e con una distanza intermicrofonica di circa 14 cm, com'è possibile anche notare in Fig. 5.

La scelta della distanza intermicrofonica è importante poiché influisce direttamente sulle caratteristiche del sistema, come descritto nel Par. 3.1. Dovendo posizionare soli due microfoni, si è scelto di distanziarli di 14 cm, poiché in questo modo si riesce ad avere una migliore risoluzione spaziale che consente di ridurre il livello di rumore in campo lontano, pur rischiando di avere un effetto di aliasing spaziale nel campo vicino. Inoltre, la distanza di 14 cm è oggi molto diffusa nei sistemi portatili.

5.2. Valutazione qualitativa del sistema di riduzione del rumore

Il miglioramento qualitativo del segnale vocale elaborato e la riduzione del rumore apportata da un sistema di beamforming vengono solitamente associate al miglioramento del rapporto segnale-rumore (SNR), definito come (Benesty et al., 2010; Uncini, 2010):

$$\text{SNR} = 10\log_{10}\left(\frac{\text{E}\left\{r_{\text{IN}}^2[n]\right\}}{\text{E}\left\{r_{\text{OUT}}^2[n]\right\} - \text{E}\left\{r_{\text{IN}}^2[n]\right\}}\right) \quad (8)$$

dove $r_{\text{IN}}[n]$ è il generico segnale di ingresso del sistema ed $r_{\text{OUT}}[n]$ rappresenta il segnale elaborato. L'operatore $\text{E}\{\cdot\}$ indica il valore atteso. Nel caso in cui volessimo misurare il rapporto segnale-rumore di ingresso al nostro sistema, SNR_{IN}, il segnale $r_{\text{IN}}[n]$ sarebbe nient'altro che il segnale desiderato $s[n]$ emesso dal parlatore, mentre il segnale $r_{\text{OUT}}[n]$ sarebbe il segnale $u_m[n]$ acquisito dal microfono. Analogamente, per ottenere un valore SNR descrittivo dell'uscita del beamformer, SNR_{OUT}, il segnale acquisito dall'interfaccia diventerebbe $r_{\text{IN}}[n]$ mentre il segnale $r_{\text{OUT}}[n]$ rappresenterebbe il segnale $e[n]$ in uscita dal beamformer. Utilizzando le misure di SNR di ingresso e di uscita è possibile definire il *guadagno d'array* o *direttività* (Uncini, 2010), come il miglioramento del rapporto segnale-rumore tra l'ingresso e l'uscita del beamformer:

$$\text{G} = \frac{\text{SNR}_{\text{OUT}}}{\text{SNR}_{\text{IN}}} \quad (9)$$

Il guadagno d'array dunque è il parametro prestazionale utilizzato per valutare la quantità di miglioramento qualitativo in dB ottenuta attraverso l'elaborazione dei segnali vocali effettuata dal sistema di beamforming. A tale riguardo consideriamo due scenari particolari del set-up descritto nel paragrafo precedente. Nel primo scenario consideriamo il segnale principale, emesso dalla sorgente desiderata, con aggiunta di rumore ambientale di fondo di vario tipo e con diversi livelli di SNR_{IN}: 0, 6 e 18 dB. Il secondo scenario è analogo al primo ma prevede l'aggiunta della sorgente interferente che riproduce un segnale vocale tratto da un notiziario. Osservando la Tabella 1 e la Tabella 2, è possibile avere una valutazione del miglioramento qualitativo, in termini di guadagno d'array, introdotto dal beamformer presentato. In particolare, si può notare dalla Tabella 1 come il sistema, in assenza di altre sorgenti interferenti, reagisca bene per rumori ambientali di fondo di tipo rosa, mentre il rumore ambientale di tipo babble risulta più difficile da cancellare.

SNR_{IN}	*Rumore rosa*	*Rumore di automobile*	*Rumore di fabbrica*	*Rumore babble*
0	5.2	3.2	3.9	3.1
6	6.9	4.5	4.8	4.7
18	7.1	5.0	5.6	5.7

Tabella 1: Guadagno d'array in dB per scenario in presenza di solo rumore di fondo.

SNR_{IN}	*Rumore rosa*	*Rumore di automobile*	*Rumore di fabbrica*	*Rumore babble*
0	1.8	1.6	1.7	1.6
6	2.7	2.3	2.4	2.6
18	4.1	4.1	3.9	3.9

Tabella 2: Guadagno d'array in presenza di rumore di fondo e sorgente interferente.

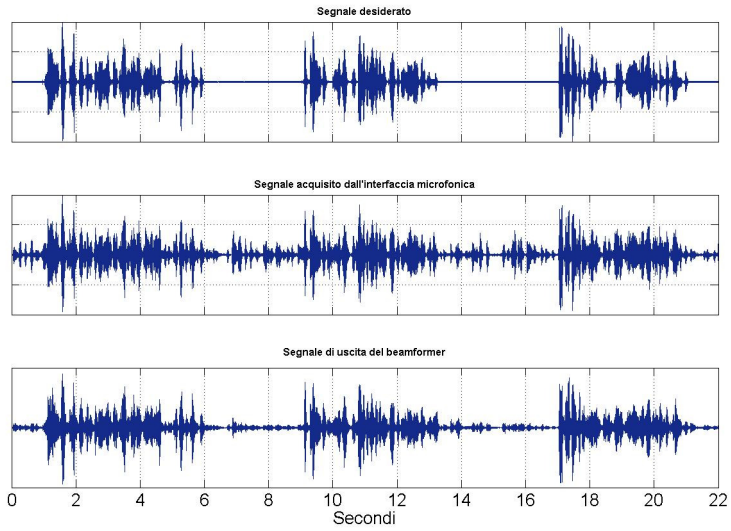

Figura 6: Segnale elaborato dal beamformer in presenza di rumore ambientale di fondo di tipo rosa e segnale interferente vocale per un SNR_{IN} = 6 dB.

Tuttavia, dalla Tabella 2 si evince che quando nell'ambiente si verifica anche la presenza di una sorgente interferente di tipo vocale, le prestazioni, come prevedibile, subiscono un peggioramento, dovuto anche alla natura tempo variante del segnale interferente. Un esempio grafico di riduzione del rumore è raffigurato in Fig. 6 ed è relativo al secondo scenario in cui è evidente la presenza della sorgente vocale interferente.

6. CONCLUSIONI

In questo lavoro è stato presentato un sistema di miglioramento qualitativo utilizzante un nuovo algoritmo adattativo, il VSS-BEAPA, il quale presenta prestazioni migliori rispetto ad altri algoritmi convenzionali in termini di velocità di convergenza e di efficienza computazionale. La bontà del sistema di beamforming è stata valutata in una configurazione piuttosto avversa, composta da una schiera microfonica "minima", ossia costituita da due soli microfoni. Tuttavia, nonostante il sistema microfonico a costo minimo utilizzato, è stato possibile constatare come il sistema di riduzione del rumore utilizzato abbia raggiunto risultati accettabili di miglioramento qualitativo del parlato per diversi tipi di rumore di fondo ambientale. Risultati migliori potranno essere raggiunti abbinando al sistema di beamforming tecniche microfoniche più adeguate consentendo così di ottenere un ulteriore miglioramento del parlato nelle comunicazioni viva voce.

BIBLIOGRAFIA

Benesty, J., Chen, J. & Huang, Y. (2010), *Microphone array signal processing*, Berlin, Heidelberg: Springer Verlag.

Bitzer, J., Simmer, K.U. & Kammeyer, K.-D. (1999), Theoretical noise reduction limits of the generalized sidelobe canceller (GSC) for speech enhancement, in *Proceedings of IEEE*

International Conference on Acoustics, Speech, and Signal Processing (ICASSP '99), Phoenix, AZ, USA, March 15-19, 2965-2968.

Campbell, D.R., Palomaki, K.J. & Brown, G.J. (2005), Roomsim, a MATLAB simulation of "shoebox" room acoustics for use in teaching and research, *Computing and Information Systems*, vol. 9, no. 3, 48-51.

Comminiello, D., Scarpiniti, M., Parisi, R. & Uncini, A. (2010), A novel affine projection algorithm for superdirective microphone array beamfomring, in *Proceedings of IEEE International Symposium on Circuits and Systems (ISCAS '10)*, Paris, France, May 30-June 2, 2127-2130.

Griffiths, L. & Jim, C. (1982), An alternative approach to linearly constrained adaptive beamforming, *IEEE Transactions on Antennas and Propagation*, vol. 30, no. 1, 27-34.

Lotter, T. & Vary, P. (2006), Dual-channel speech enhancement by superdirective beamforming, *EURASIP Journal on Applied Signal Processing*, vol. 2006, no. 1, 1-14.

Luo, F.-L. & Uvacek, B. (2002), The combination of binaural processing with adaptive beamforming for dual microphones in speech communications, in *Proceedings of IEEE Circuits and Systems for Communications (ICCSC '02)*, St. Petersburg, Russia, June 26-28, 436-439.

Marro, C., Mahieux, Y. & Simmer, K.U. (1998), Analysis of noise reduction and dereverberation techniques based on microphone arrays with postfiltering, *IEEE Transactions on Speech and Audio Processing*, 6, 240-259.

McCowan, I.A. & Bourlard, H. (2002), Microphone array post-filter for diffuse noise field, in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '02)*, Orlando, FL, USA, May 13-17, 905-908.

Ozeki, K. & Umeda, T. (1984), An adaptive filtering algorithm using an orthogonal projection to an affine subspace and its properties, *Electronics & Communications in Japan*, vol. 67-A, 19-27.

Paleologu, C., Ciochina, S. & Benesty, J. (2008), Variable step-size NLMS algorithm for under-modeling acoustic echo cancellation, *IEEE Signal Processing Letters*, 15, 5-8.

Sayed, A.H. (2008), *Adaptive filters*, Hoboken, NJ: John Wiley & Sons, Inc.

Stern, R.M., Gouvea, E., Chanwoo, K., Kumar, K. & Park, H.-M. (2008), Binaural and multiple-microphone signal processing motivated by auditory perception, in *Proceedings of Hands-Free Speech Communication and Microphone Arrays (HSCMA 2008)*, Trento, Italy, May 6-8, 98-103.

Tanaka, M., Makino, S. & Kojima, J. (1999), A block exact fast affine projection algorithm, *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 1, 79-86.

Uncini, A. (2010), *Elaborazione adattativa dei segnali*, Roma: Aracne Editrice S.R.L.

Zheng, Y.R. & Goubran, R.A. (2000), Adaptive beamforming using affine projection algorithms, in *Proceedings of the 5th International Conference on Signal Processing (WCCC-ICSP '00)*, Beijing, China, August 21-25, 1929-1932.