Audio signal digital processing method and system thereof

MINISTERO DELLO SVILUPPO ECONOMICO
DIREZIONE GENERALE PER LA LOTTA ALLA CONTRAFFAZIONE
UFFICIO ITALIANO BREVETTI E MARCHI

DOMANDA DI INVENZIONE NUMERO	**102017000073663**
Data Deposito	**30/06/2017**
Data Pubblicazione	**30/12/2018**

Classifiche IPC

Titolo

Audio signal digital processing method and system thereof

Metodo di processo digitale di un segnale audio e relativo sistema
Audio signal digital processing method and system thereof

DESCRIZIONE

CAMPO TECNICO

La presente invenzione si riferisce a un metodo di processo del segnale audio e a un relativo sistema, in particolare per l'uso in un ambiente di lavoro come un open space o un ambiente di lavoro per colletti bianchi o simili.

STATO DELL'ARTE

E' noto suddividere un open space tramite cubicoli per definire una stazione di lavoro per un colletto bianco. Un cubicolo non offre un assorbimento audio efficace all'interno dell'open space. Lo stesso si applica ad altre tecniche di assorbimento audio, incluse quelle che prevedono pannelli di assorbimento audio o simili all'interno dell'open space. Le soluzioni attuali causano un certo livello di inquinamento rumoroso dovuto all'interazione fra persone (chiaccherate, telefonate, etc.) o al sottofondo. Ciò causa la perdita di concentrazione, un aumento di stress e conseguentemente una produttività ridotta per i lavoratori. Inoltre, tale rumore influenza la qualità della comunicazione vocale con altre persone perché in un ambiente rumoroso e spesso più difficile e più stressante parlare con colleghi.

E' inoltre noto fornire dispositivi passivi di noise cancellation come tappi per le orecchie.

Al contempo, è noto fornire dispositivi di active noise cancellation per ridurre audio (rumore) indesiderato tramite la generazione di un secondo audio

specificamente progettato per interferire e cancellare l'audio indesiderato.

SCOPI E RIASSUNTO DELL'INVENZIONE

Lo scopo della presente invenzione è di processare l'audio di un ambiente, in particolare in un open space o altro luogo delimitato o di lavoro, per eliminare audio indesiderato per aumentare la concentrazione dei lavoratori nell'open space o sul luogo di lavoro.

Lo scopo della presente invenzione è raggiunto da un metodo di processo audio e un sistema in grado di separare i segnali audio di ciascuna sorgente audio puntuale da un segnale audio misto di un open space o di un luogo di lavoro, di separare un numero di sorgenti audio puntuali, di riconoscere chi parla, e di riprodurre, tramite un altoparlante indossabile in-ear, on-ear, o around-the-ear solamente quelle tracce audio che appartengono all'audio dell'ambiente e che si abbinano a una lista di sorgenti audio puntuali accettate, includente sorgenti audio puntuali umane.

In questo modo, un gruppo di lavoratori può aggiungere l'un l'altro nella lista di sorgenti audio puntuali accettate e tenere una riunione di persona in un ambiente di lavoro affollato e le voci rilevanti saranno enfatizzati tramite l'altoparlante all'orecchio rispetto alle voci provenienti da sorgenti audio non accettate. Infatti, durante questa riunione, l'altoparlante indossabile di ciascun partecipante riprodurre le voci degli altri partecipanti e non riprodurla altri audio o rumori.

Altri vantaggi e caratteristiche della presente invenzione sono discussi nella descrizione e citati nelle rivendicazioni dipendenti.

BREVE DESCRIZIONE DEI DISEGNI

L'invenzione viene descritta nel seguito sulla base di esempi non limitativi

illustrati a scopi esplicativi nel disegno annesso, che si riferisce a uno schizzo di un sistema digitale di processo audio

DESCRIZIONE DETTAGLIATA DELL'INVENZIONE

La figura mostra, nel complesso, un ambiente, in particolare un ambiente di lavoro, dove utenti, in particolare colleghi, stanno parlando simultaneamente.

L'utente A e un utente B sono in riunione e un utente C sta parlando ad un telefono cellulare con una persona non illustrata e non sta interagendo con gli utenti A, B. In tale condizione, i discorsi degli utenti A, B sono disturbati dal discorso dell'utente C e viceversa. L'ambiente comprende inoltre un allarme AB includente un emettitore di un segnale audio d'allarme (non illustrato), per esempio un allarme antincendio, e una rete di antenne N per uno scambio senza fili di dati con una larghezza di banda adatta, cioè maggiore il numero di antenne per unità di area dell'ambiente, maggiore larghezza di banda per lo scambio dati. Le antenne N sono, per esempio, antenne wi-fi® o antenne Bluetooth®. È importante notare che, per incrementare la larghezza di banda per lo scambio dati, sono coinvolti anche gain, potenza di trasmissione/ricezione e potenza di calcolo.

Ciascuna sorgente audio citata sopra, genera durante la normale attività nell'ambiente una relativa traccia o canale audio che si mescola con altre tracce audio in un segnale audio misto dell'ambiente includente i discorsi simultanei de, ad esempio, gli utenti A, B, C.

Un metodo secondo l'invenzione comprende la fase di fornire una lista di una o più sorgenti audio accettate, includente una sorgente audio puntuale umana o di discorso. Tale fase può essere implementata dagli utenti tramite un'interfaccia utente, come uno smartphone o un computer da scrivania, in modo

da inserire gli utenti ammessi alla riunione.

Preferibilmente, la lista di sorgenti audio accettate è generata o completata sulla base di dati ricevuti da un sistema di gestione del tempo del calendario. In particolare, i partecipanti alla riunione sono invitati e confermati tramite il sistema di gestione del tempo e del calendario e la lista di sorgenti accettate è generata, preferibilmente generata automaticamente, includendo i partecipanti invitati e/o confermati memorizzati nel sistema di gestione del tempo e del calendario. È possibile generare una nuova lista di sorgenti sonore accettate per ciascuna riunione prevista o di modificare la lista precedente tramite l'interfaccia utente.

I partecipanti alla riunione sono presenti in un ufficio open space o ambiente simile comprendente una stanza condivisa con altri colleghi che non partecipano alla riunione. La reta di antenne N copre lo spazio dell'ambiente in modo da fornire una banda adeguata per lo scambio dati.

In un'ulteriore fase, il segnale audio misto di discussione simultanea dall'ambiente dove gli utenti A, B, C stanno parlando è acquisito tramite un'unità di microfono M. L'unità di microfono comprende uno o più microfoni localizzati all'interno dell'ambiente. I microfoni possono avere posizioni fisse all'interno dell'ambiente oppure possono essere microfoni portatili, come quelli previsti a bordo di cuffie e/o di dispositivi portatili intelligenti come smartphone, tablet (normalmente definiti come dispositivi personali intelligenti portatili) e computer portatili. La posizione e la copertura dell'unità di microfono delimita l'area come l'open space, a cui e applicato il metodo.

Il segnale audio misto di discorsi simultanei è processato da un'unità di processo digitale D in modo da isolare la traccia audio proveniente da ciascuna

sorgente audio, tramite un algoritmo di Blind Source Separation (BSS), come un algoritmo Barra-Spence. Per esempio, un algoritmo BSS è basato su tecniche multicanale e su tecniche Time Direction of Arrival (TDOA).

L'unità di processo digitale D è o un'unità a processore unico o un'architettura multi-processore, includente un'unità multi-processore cloud computing per ridurre il tempo elaborazione dati.

L'unità di processo digitale D riconosce inoltre la sorgente audio puntuale e le tracce provenienti da tali sorgenti audio puntuali, includenti le tracce audio degli utenti A, B, C. Ciò è eseguito per esempio tramite un algoritmo di Source Recognition.

E' possibile eseguire la separazione di ciascuna traccia appartenente alla singola sorgente audio puntuale e, successivamente, il riconoscimento di ciascuna traccia in modo da assegnare ciascuna traccia alla relativa sorgente audio puntuale. Per esempio, una traccia mista comprendente voci degli utenti A e B è inizialmente elaborata per isolare due tracce e successivamente, l'unità di processo digitale D riconosce quale traccia appartiene all'utente A e quale traccia appartiene all'utente B.

In modo da risparmiare tempo e ridurre la latenza del metodo, è preferibile includere una fase di fornire una libreria di dati audio di apprendimento delle sorgenti audio puntuali. Ciò è per esempio implementato tramite il processo di una traccia di discorso singola di ciascun collega e assegnazione dei dati rilevanti al relativo collega. Tali dati nella libreria fungono come rispettivi dizionari durante la separazione e il riconoscimento.

Secondo una forma di realizzazione non limitativa della presente invenzione, la libreria comprende inoltre i dati da una sorgente audio puntuale

non umana o non di discorso, come allarmi, campanelli della porta, etc. che sono in una posizione fissa all'interno dell'ambiente delimitato, come un allarme AB, e che possono entrare o uscire dall'ambiente delimitato, come utenti umani. Possono essere considerate anche suonerie di telefoni cellulari.

Un esempio di elaborazione di tracce audio in modo da estrarre opportuni dati di dizionario per la libreria è la preparazione di un impronta digitale audio biometrica tramite quantizzazione o fattorizzazione vettoriale/matriciale delle tracce audio di apprendimento, in cui ciascuna sorgente audio puntuale è processata in modo da generare un vettore/matrice di identificazione, cioè i dati di apprendimento, inclusi nella libreria. Durante la separazione e/o il riconoscimento, l'unità di processo digitale D elabora quindi l'audio ambientale misto per identificare il vettore/ matrice di identificazione della libreria. In modo da aumentare l'attendibilità del metodo, è preferibile che la libreria includa dati di apprendimento provenienti da un elevato numero di sorgenti audio puntuali differenti correlati all'ambiente delimitato, come laboratori e sorgenti sonore fisse dell'open space. In particolare, quando il metodo è implementato in un ambiente di lavoro, dati di apprendimento di tutti i colleghi e altre sorgenti sonore puntuali non umane presenti nel luogo di lavoro, cioè l'open space, sono processate e aggiunte alla libreria, includendo dati processati da tracce audio di allarmi, ad esempio un allarme incendio da un allarme AB. Quando il metodo è applicato in un ambiente delimitato, come un open space, è preferibile che le tracce della maggioranza o di tutte le sorgenti sonore puntuali all'interno dell'ambiente siano processate in modo da estrarre dati di apprendimento da aggiungere alla libreria.

In modo da ridurre il tempo di elaborazione e la latenza, è preferibile approcciare sia la separazione tramite l'algoritmo BSS che il riconoscimento

tramite l'algoritmo Source Recognition sulla base di dati di apprendimento della libreria, in particolare le sorgenti audio puntuali sono caratterizzate durante la separazione delle tracce tramite i dizionari raccolti nella libreria.

Quanto sopra è preferibilmente implementato tramite un algoritmo di Non-negative Matrix Factorization. I dettagli sono spiegati nel seguito e, in aggiunta, in altri paragrafi tratti da J. Zegers, H. Van Hamme in 'Joint Audio Source Separation and Speaker Recognition', April 29th, 2016 - arXiv:1604.08852v1:

> La Non-negative matrix factorization è un metodo di fattorizzazione che approssima una matrice non negativa $\mathbf{X} \in \mathbb{R}_+^{F \times N}$ utilizzando una matrice dizionario non negativa $\mathbf{T} \in \mathbb{R}_+^{F \times K}$ e una matrice di attivazione non negativa $\mathbf{V} \in \mathbb{R}_+^{K \times N}$, in modo che $\mathbf{X} \approx \hat{\mathbf{X}} \triangleq \mathbf{TV}$. Nella nostra applicazione $\mathbf{X} = |\tilde{\mathbf{X}}|^{\cdot 2}$ è uno spettrogramma di potenza di discorso, dove $\tilde{\mathbf{X}}$ è la trasformata di Fourier short time (STFT) a valori complessi del segnale audio, $|\cdot|$ è il valore assoluto e $\cdot^2$ è il quadrato d'insieme dell'elemento. $\mathbf{X}$ è la matrice con *F* campioni di spettro e *N* intervalli di tempo. Il NFM tenta di catturare gli andamenti più frequenti del discorso in vettori di base dimensionale *K F* che formano un dizionario **T** per il discorso. La matrice **V** contiene i coefficienti della combinazione lineare e perciò indica come il vettore di base *k* è attivato nell'intervallo di tempo *n*. In genere, $K < min(F, N)$ in modo che il NMF sia un'operazione di riduzione del rango. Una misura di discrepanza è scelta fra la **X** originale e la $\hat{\mathbf{X}}$ ricostruita e può essere minimizzata tramite l'individuazione di dizionari e attivazioni ottimali. La distanzia euclidea (EU), la divergenza di Kullback-Leibler (KL) e la

divergenza di Itakura-Saito (IS) sono misure ben note. In questo articolo sarà usata la divergenza IS.

$$d_{IS}(x_{fn}, \hat{x}_{fn}) = \frac{x_{fn}}{\hat{x}_{fn}} - log\left(\frac{x_{fn}}{\hat{x}_{fn}}\right) - 1 \qquad (1)$$

Per minimizzare questa divergenza, sono state derivate garanzie di convergenza formule iterative moltiplicative con garanzie di convergenza [14].

$$t_{fk} \leftarrow t_{fk} \sqrt{\frac{\sum_n \frac{x_{fn}}{\hat{x}_{fn}} \frac{v_{kn}}{\hat{x}_{fn}}}{\sum_n \frac{v_{kn}}{\hat{x}_{fn}}}} \qquad (2)$$

$$v_{kn} \leftarrow v_{kn} \sqrt{\frac{\sum_f \frac{x_{fn}}{\hat{x}_{fn}} \frac{t_{fk}}{\hat{x}_{fn}}}{\sum_f \frac{t_{fk}}{\hat{x}_{fn}}}} \qquad (3)$$

dove i sotto-indici si riferiscono all'elemento corrispondente nella matrice. Per evitare ambiguità di scale le colonne di **T** devono essere normalizzate: $t_{fk} \leftarrow t_{fk} / \sum_{f^*} t_{f^*k}$.

L'uso della NMF in applicazioni SR per un singolo discorso è semplice. Nella fase di addestramento, i dati di addestramento del j-esimo utente che parla $\mathbf{X}^j_{train}$ è fattorizzato usando le equazioni 2 e 3. I dizionari ottenuti $\mathbf{T}^j$ per ciascuno dei J utenti che parlano, sono ipotizzati essere dipendenti dagli utenti che parlano e sono raccolti nella libreria $\mathbf{T}_{tot} = [\mathbf{T}^1, \mathbf{T}^2, \ldots, \mathbf{T}^J]$.

Durante la verifica l'identità di un utente che parla s deve essere trovate in una $\mathbf{X}^s_{test}$ precedentemente non vista. La NFM è applicata con una libreria fissa $\mathbf{T}_{tot}$ e le attivazioni $\mathbf{V}^s_{tot,test}$ sono trovate iterativamente usando l'equazione 3. La matrice di attivazione indica quantitativamente l'attivazione di ciascun vettore di base per ciascun

utente bersaglio che parla in ciascun intervallo di tempo.

L'attività combinata di tutti i vettori di base in un dizionario di un utente bersaglio che parla è una misura dell'attività dell'utente bersaglio che parla nel segmento di test. È possibile includere vincoli Group Sparcity (GS-NMF) sulle attivazioni $\mathbf{V}_{tot,test}^{s}$ per mettere in pratica soluzioni in cui è improbabile che i vettori di base da utenti bersaglio differenti che parlano siano attivi nello stesso intervallo di tempo [15] [16]. Un modo semplice per stimare l'identità di un utente che parla è di determinare l'utente bersaglio che parla per cui la somma delle attivazioni, su tutti i suoi vettori di base e su tutti gli intervalli di tempo, è massima. Questo modo di classificare può essere visto come una stima della probabilità per intervallo dove la stima finale è una media pesata di tutti gli intervalli, dando più peso agli intervalli con una maggiore attivazione o con più energia.

$$\hat{ID}_s = \arg\max_j \sum_{k\in\kappa^j} \sum_n v_{tot,kn}^{s} \qquad (4)$$

dove κ^j sono gli indici dei vettori di base appartenenti al dizionario dell'utente che parla j-esimo. E' possibile eseguire una classificazione più avanzata delle attivazioni all'identità di un utente che parla utilizzando, per esempio a vettore di supporto.

Oltre ad applicazioni SR, la NMF ha anche dimostrato buoni risultati in problemi di source separation. Quando i differenti utenti che parlano sono appresi su dati di apprendimento di un singolo discorso, la procedura è molto simile a quella della SR. Tuttavia, nella

SS, i dati di test $\mathbf{X}_{test}$ contengono un discorso di sorgenti multiple che parlano simultaneamente. Il compito non è determinare l'identità dell'utente che parla, ma il segnale sorgente originale di ciascun utente che parla.

Dopo aver appreso $\mathbf{T}_{tot}$ nella fase di apprendimento, $\mathbf{V}_{tot,test}$ è calcolata nello stesso modo della sezione 2.2. Usando un filtraggio Wiener e la fase delle osservazioni, il segnale sorgente originale $\hat{y}^s_{fn}$ può essere stimato [8].

$$\hat{y}^s_{fn} = \left(\frac{\sum_{k \in \kappa^s} t_{fk} v_{kn}}{\sum_{s^*} \sum_{k \in \kappa^{s^*}} t_{fk} v_{kn}} |\tilde{x}_{fn}| \right) arg(\tilde{x}_{fn}) \qquad (5)$$

dove κ^ssono gli indici dei vettori di base appartenenti al dizionario dell'utente che parla *s* e $arg(\tilde{x}_{fn})$ denota la fase di $\tilde{x}_{fn}$.

In molte situazioni, tuttavia, non c'è possibilità di separazioni di sorgente supervisionata. Nella blind source separation (BSS) non è disponibile $\mathbf{X}^s_{train}$ per apprendere la libreria $\mathbf{T}_{tot}$. Invece la libreria sarà creata durante la separazione stessa. Normalmente in questi casi si ricorre a tecniche multicanale, in cui tecniche Time Direction Of Arrival (TDOA) possono essere usate per supportare la separazione della sorgente. La matrice di mixing $\mathbf{M} \in \mathbb{C}^{F \times I \times S}$ è considerata statica e quindi indipendente da *n*.

$$\tilde{x}_{i,fn} = \sum_s m_{is,f} y_{s,fn} \qquad (6)$$

dove I è la quantità di microfoni e $m_{is,f}$ indica la rappresentazione nel

dominio della frequenza della risposta all'impulso della stanza (RIR) tra l'utente che parla *s* e il microfono *i* per il campione di spettro *f*. $\tilde{x}_i$ è lo spettrogramma STFT ricevuto nel microfono *i* e y_s è lo spettrogramma STFT del segnale originale dell'utente che parla *s*. A causa dell'ambiguità di scala dell'equazione 6, possono essere stimate solo le RIR relative fra i microfoni. La notazione per i segnali combinati dei microfoni è la seguente.

$$\tilde{\mathbf{x}}_{fn} = [\tilde{x}_{1,fn}, \tilde{x}_{2,fn}, \ldots, \tilde{x}_{I,fn}] \qquad (7)$$

Sawada et a. hanno proposto una divergenza IS multicanale [9].

$$D_{IS}(\mathbf{X}, \{\mathbf{T}, \mathbf{V}, \mathbf{H}, \mathbf{Z}\}) = \sum_f \sum_n d_{IS}(X_{fk}, \hat{X}_{fk})$$
$$d_{IS}(X_{fk}, \hat{X}_{fk}) = tr(X_{fk}\hat{X}_{fk}) - logdet(X_{fk}\hat{X}_{fk}) \qquad (8)$$

dove $tr(.)$ è la traccia di una matrice, $logdet(.)$ è il logaritmo naturale del determinante di una matrice, $X_{fn} = \tilde{\mathbf{x}}_{fn}\tilde{\mathbf{x}}_{fn}^H$ con $.^H$ la trasposta Hermitiana di una matrice e $\hat{X}_{fn} = \sum_k \left(\sum_s H_{fs} z_{sk}\right) t_{fk} v_{kn}$. t_{fk} e v_{kn} hanno le stesse interpretazioni che nel NMF a canale singolo. z_{sk} è un indicatore latente di utente che parla per indicare la certezza che il vettore *k* appartiene al dizionario dell'utente che parla *s* rispettando le condizioni $z_{sk} \geq 0$ e $\sum_s z_{sk} = 1$. H_{fs} è una matrice Hermitiana semi-definita positiva con, sulle proprie diagonali, il gain di potenza dell'utente che parla *s* al campione di frequenza *f* per ciascun microfono. Gli elementi fuori-diagonale includono le differenze di fasi

tra i microfoni e pertato contengono informazioni spaziali dell'utente che parla. Formule di aggiornamento moltiplicative sono state trovate in [9, eq. 42-47] per minimizzare la divergenza dell'equazione 8. I segnali seprati sono successivamente ottenuti tramite filtraggio Wiener.

$$\hat{y}_{fn}^{s} = \left(\sum_{k} z_{sk} t_{fk} v_{kn} \right) H_{fs} \hat{X}_{fn}^{-1} \bar{\mathbf{x}}_{fn} \qquad (9)$$

Quando viene eseguita una speaker recognition in scenari di discorsi simultanei, si può adottare un approccio sequenziale. Prima applicare la blind source separation per ottenere segmenti multipli, presumibilmente di un discorso singolo, da un discorso simultanea. Procedere come se questi segmenti non contenessero alcuna conversazione incrociata e applicare una speak recognition come nella sezione 2.2. Tuttavia in questo articolo è scelto un approccio congiunto, in cui gli utenti che parlano sono caratterizzati tramite dizionari mentre le sorgenti sono separate. Durante l'apprendimento, la source separation è eseguita come spiegato nella sezione 2.3. Il vettore di base *k* è quindi assegnato all'utente che parla *s* per cui z_{sk} è massimo. I dizionari sono raccolti nella libreria $\mathbf{T}_{\mathbf{tot}}$. Durante il test, un algoritmo di source separation simile è usato ma $\mathbf{T}_{\mathbf{tot}}$ rimane fisso. Dal momento che ogni vettore di base è contenuto in un dizionario, il significato della variabile **Z** è cambiato. Esso mappa ora un dizionario completo *j*, e la sua corrispondente identità dell'utente bersaglio parla, rispetto a un utente di test *s* che parla. Un nuovo indicatore variabile

c_{jk} è introdotto per assegnare a un vettore di base *k* a un dizionario *j* se $c_{jk} = 1$ alle condizioni $c_{j*k} \geq 0$ e $\sum_{j*} c_{j*k} = 1$.

La variabile $\hat{X}_{fn}$ è pertanto riformulata come segue.

$$\hat{X}_{fn} = \sum_k \sum_j \sum_s H_{fs} z_{sj} c_{jk} t_{fk} v_{kn} \qquad (10)$$

Può essere facilmente dimostrato che le seguenti formule di aggiornamento si possono pertanto estendere [9, eq. 42-47].

$$t_{fk} \leftarrow t_{fk} \sqrt{\frac{\sum_j c_{jk} \sum_s z_{sj} \sum_n v_{kn} tr(\hat{X}_{fn}^{-1} X_{fn} \hat{X}_{fn}^{-1} H_{fs})}{\sum_j c_{jk} \sum_s z_{sj} \sum_n v_{kn} tr(\hat{X}_{fn}^{-1} H_{fs})}} \qquad (11)$$

$$v_{kn} \leftarrow v_{kn} \sqrt{\frac{\sum_j c_{jk} \sum_s z_{sj} \sum_f t_{fk} tr(\hat{X}_{fn}^{-1} X_{fn} \hat{X}_{fn}^{-1} H_{fs})}{\sum_j c_{jk} \sum_s z_{sj} \sum_f t_{fk} tr(\hat{X}_{fn}^{-1} H_{fs})}} \qquad (12)$$

$$z_{sj} \leftarrow z_{sj} \sqrt{\frac{\sum_k c_{jk} \sum_f \sum_n t_{fk} v_{kn} tr(\hat{X}_{fn}^{-1} X_{fn} \hat{X}_{fn}^{-1} H_{fs})}{\sum_k c_{jk} \sum_f \sum_n t_{fk} v_{kn} tr(\hat{X}_{fn}^{-1} H_{fs})}} \qquad (13)$$

$$c_{jk} \leftarrow c_{jk} \sqrt{\frac{\sum_s z_{sj} \sum_f \sum_n t_{fk} v_{kn} tr(\hat{X}_{fn}^{-1} X_{fn} \hat{X}_{fn}^{-1} H_{fs})}{\sum_s z_{sj} \sum_f \sum_n t_{fk} v_{kn} tr(\hat{X}_{fn}^{-1} H_{fs})}}$$

Per aggiornare H_{fs} è risolta un'equazione algebrica di Riccati.

$$H_{fs} A H_{fs} = B \qquad (15)$$

$$A = \sum_j \sum_k c_{jk} z_{sj} t_{fk} \sum_n v_{kn} \hat{X}_{fn}^{-1} \qquad (16)$$

$$B = H'_{fs} \left(\sum_j \sum_k c_{jk} z_{sj} t_{fk} \sum_n v_{kn} \hat{X}_{fn}^{-1} X_{fn} \hat{X}_{fn}^{-1} \right) H'_{fs} \qquad (17)$$

dove H'_{fs} è la H_{fs} dell'aggiornamento precedente. Per evitare ambiguità di scala, dovrebbero seguire queste normalizzazioni:

$H_{fs} \leftarrow H_{fs}/tr(H_{fs}), t_{fk} \leftarrow t_{fk}/\sum_{f^*} t_{f^*k}, z_{sj} \leftarrow z_{sj}/\sum_{j^*} z_{sj^*}$ e $c_{jk} \leftarrow c_{jk}/\sum_{j^*} c_{j^*k}$

Nella fase di test, un vettore di base è mantenuto fisso a un dizionario. Pertanto, $c_{jk} = 1$ solo se il vettore di base *k* appartiene al dizionario *j*, altrimenti $c_{jk} = 0$.

Usando l'equazione 14 e la normalizzazione, si può vedere che i valori per c_{jk} sono pertanto fissi per l'intero processo iterativo.

L'ID stimato per l'utente che parla *s* è *j* tale che z_{sj} è massimo.

$$\hat{ID}_s = \arg\max_j z_{sj} \qquad (18)$$

Tramite H_{fs} e z_{sj}, la speaker recognition può essere interpretata come assegnare il dizionario *j*, e quindi l'identità dell'utente bersaglio che parla associata al dizionario *j*, alla posizione dell'utente che parla *s* nell'affermazione di test. Occorre notare che *S*, la quantità di utenti che parlano nella conversazione mista, può essere uguale o inferiore a *J*, la quantità di utenti che parlano nella libreria. Non tutti gli utenti che parlano devono comparire nella conversazione mista.

La precedente discussione è limitata alla speaker recognition, cioè a sorgenti audio puntuali di discorso o umane, ma è possibile estendere l'esempio a

qualsiasi sorgente audio puntuale, includendo sorgenti audio non umane e/o sorgenti audio non di discorso, in modo da avere una panoramica e controllare un ambiente delimitato all'interno di un edificio, come un open space.

Preferibilmente, gli allarmi sono originariamente inclusi nella lista di sorgenti audio puntuali.

Secondo la presente invenzione, dopo che il segnale audio misto di discorso è processato, il metodo comprende la fase di riprodurre tramite un altoparlante da orecchio indossabile (in-ear, on-ear o around-the-ear) S usato dagli utenti che parlano presenti nella lista delle sorgenti audio puntuali accettate, ciascuna traccia audio di discorso della sorgente audio puntuale umana dell'ambiente che si abbina con quelle della lista di sorgenti audio puntuali accettate. In particolare, ammesso che gli utenti A, B, C abbiano i rispettivi dati di discorso audio processati e memorizzati nella libreria come dati di dizionario/apprendimento, gli altoparlanti da orecchio degli utenti A, B, che sono in riunione e sono nella lista delle sorgenti sonore puntuali accettate per quella riunione, riprodurranno le tracce audio provenienti dagli utenti A, B e non la traccia audio dell'utente C.

Secondo una forma di realizzazione preferita, il segnale audio misto di discorso simultaneo è processato prima della separazione e del riconoscimento in modo da implementare un'active noise cancelling del rumore di fondo o simili. Ciò è per esempio implementato tramite filtraggio.

I vantaggi dell'invenzione sono i seguenti.

Tutte le tecnologie esistenti consentono agli utenti di diminuire il rumore indesiderato, ma non consentono di selezionare fra alcuni suoni specifici desiderati o accettati e altri segnali sonori, identificati come "rumore". Inoltre, non forniscono la caratteristica di selezionare l'utente che parla, cioè la possibilità

di riconoscere le voci fra il rumore e selezionare all' occorrenza solo alcune voci specifiche da essere udibili dagli utenti. Infatti, con le soluzioni correnti non è possibile sentire solo alcune voci umane selezionate e scartare voce indesiderate e il rumore di fondo. Pertanto, la soluzione proposta si prefigge di proteggere il lavoratore dall'inquinamento acustico ma non lo isola completamente (acusticamente parlando) dall'ambiente e dalle persone con cui l'utente può voler interagire. Infatti, l'altoparlante da orecchio fornisce una porzione di smorzamento passivo del rumore ma tale smorzamento non è assoluto. La soluzione proposta incrementerà la concentrazione e le prestazioni dell'utente quando al lavoro senza compromettere la sua possibilità di interagire, migliorando la qualità della comunicazione verbale in ambienti di lavoro con inquinamento acustico.

Secondo l'invenzione è possibile limitare la latenza tra la fase di acquisizione e la fase di riproduzione a meno di o pari a 100 millisecondi. Tale soglia è considerata tale che l'utente non si accorga del ritardo temporale tra il linguaggio parlato e la riproduzione tramite gli altoparlanti da orecchio S. Ciò assicura un'esperienza discreta.

La presente invenzione si adatta al meglio per gli ambienti di lavoro dove colleghi tengono riunioni e sono spesso disposti in un open space. Tuttavia, la presente invenzione è anche applicabile in ambienti simili, come ad esempio club, aree comuni per studenti o simili, dove persone si raggruppano e parlano in gruppi simultanei all'interno della medesima grande sala ed è preferibile evitare che le voci dei gruppi si sovrappongano l’un l'altra.

Bibliografia:

[8] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1066–1074, 2007.

[9] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Multichannel extensions of non-negative matrix factorization with complex-valued data," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 5, pp. 971–982, 2013.

[14] M. Nakano, H. Kameoka, J. Le Roux, Y. Kitano, N. Ono, and S. Sagayama, "Convergence-guaranteed multiplicative algorithms for nonnegative matrix factorization with β-divergence," *In Proceedings of the 2010 IEEE International Workshop on Machine Learning for Signal Processing (MLSP), Kittila, Finland, 29 August – 1 September 2010*, pp. 283–288.

[15] A. Lefèvre, F. Bach, and C. Févotte, "Itakura-saito nonnegative matrix factorization with group sparsity," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 21–24.

[16] A. Hurmalainen, R. Saeidi, and T. Virtanen, "Group sparsity for speaker identity discrimination in factorisation-based speech recognition." in *INTERSPEECH*, 2012, pp. 2138–2141.

RIVENDICAZIONI

1. Metodo di processo digitale audio comprendente le fasi di:

- fornire una lista di sorgenti audio puntuali accettate, includente una sorgente audio puntuale di discorso o umana;

- acquisire in un ambiente un segnale audio di discorso simultaneo misto da una o più sorgenti audio puntuali dell'ambiente, includente una sorgente audio puntuale dell'ambiente umana o di discorso, tramite un'unità di microfono;

- processare il segnale audio di discorso simultaneo misto in modo da:

separare le tracce audio di ciascuna delle sorgenti audio puntuali dell'ambiente tramite un algoritmo Blind Source Separation

riconoscere le tracce tramite un algoritmo di Source Recognition

Riprodurre tramite un altoparlante indossabile all'orecchio (in-ear, on-ear or around-the-ear), ciascuna traccia di discorso della sorgente audio puntuale umana dell'ambiente che si abbina con quelle della lista di sorgenti audio puntuali accettate.

2. Metodo secondo la rivendicazione 1, comprendente inoltre la seguente fase prima della fase di processare:

- fornire una libreria di dati di apprendimento per identificare le sorgenti audio puntuali;

in cui la lista delle sorgenti audio puntuali accettate è selezionata dalla libreria; e

in cui sia la separazione che il riconoscimento nella fase di processo sono basati su dati codificati nella libreria.

3. Metodo secondo una rivendicazione 2, in cui sia la separazione che

il riconoscimento dello stadio di processo sono basati su un algoritmo si Non-negative Matrix Factorization in qualità di algoritmo combinato di Source Separation and Source Recognition

4. Metodo secondo la rivendicazione 3, in cui la latenza fra la fase di acquisire e la fase di riprodurre è minore o uguale a 100 millisecondi.

5. Metodo secondo una qualsiasi delle rivendicazioni precedenti, in cui è applicato un algoritmo di Active Noise Cancellation e la fase di riprodurre comprende una riproduzione di un segnale audio di cancellamento per interferire con le tracce di una o più sorgenti audio puntuali escluse dalla lista delle sorgenti audio puntuali accettate.

6. Metodo secondo una qualsiasi delle rivendicazioni precedenti, in cui la fase di fornire la lista è eseguita tramite un'interfaccia utente.

7. Metodo secondo la rivendicazione 6, in cui la fase di fornire la lista è basata su dati relativi alle sorgenti audio puntuali umane processate da un sistema di gestione del calendario per la partecipazione a una riunione.

8. Metodo Secondo una qualsiasi delle rivendicazioni precedenti, in cui una sorgente sonora di allarme dell'ambiente è originariamente inclusa nella lista di sorgenti audio puntuali accettate.

9. Sistema di processo audio comprendente:

- un dispositivo di memorizzazione per memorizzare una lista di sorgenti audio puntuali accettate, includente una sorgente audio puntuale di discorso o umana;

- un'unità di microfono per acquisire in un ambiente un segnale audio di discorso simultaneo misto da una o più sorgenti audio puntuali

dell'ambiente, includente una sorgente audio puntuale dell'ambiente umana o di discorso;

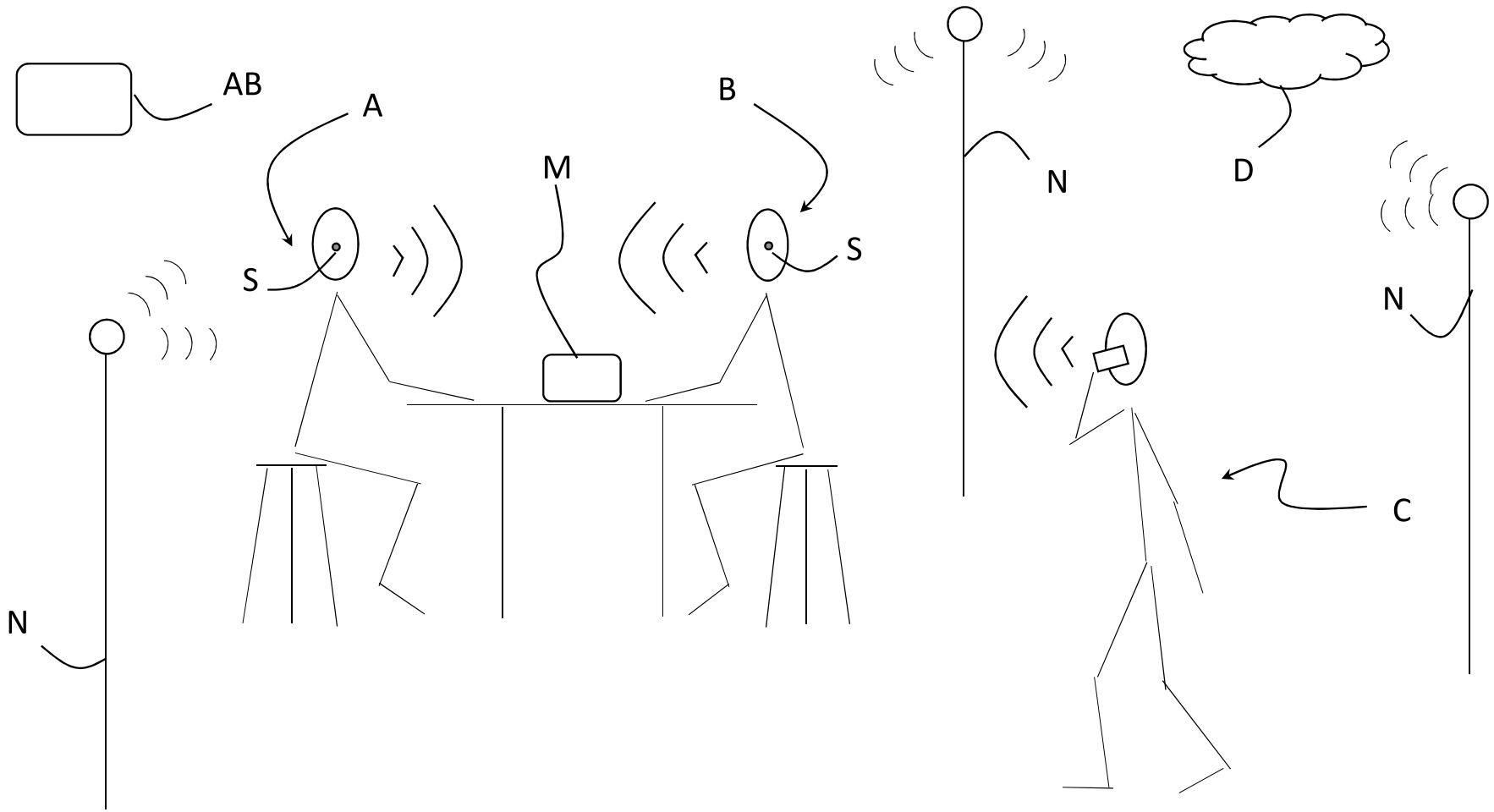
- un dispositivo di processo per processare il segnale audio di discorso simultaneo misto in modo da:

separare le tracce audio di ciascuna delle sorgenti audio puntuali dell'ambiente tramite un algoritmo Blind Source Separation

riconoscere le tracce tramite un algoritmo di Source Recognition

- un altoparlante indossabile all'orecchio (in-ear, on-ear or around-the-ear), per riprodurre ciascuna traccia di discorso della sorgente audio puntuale umana dell'ambiente che si abbina con quelle della lista di sorgenti audio puntuali accettate.

10. Sistema secondo rivendicazione 9, caratterizzato dal fatto di comprendere una pluralità di antenne per interconnettere il microfono e/o l'altoparlante da orecchio al dispositivo di processo e scambiare senza fili dati relativi alla lista e/o al segnale audio di discorso simultaneo misto e/o alle tracce audio di discorso.


AB
A
B
M
S
S
N
N
D
N
C