Modellazione fisica della glottide e inversione acustico-articolatoria

E. Marchetto, F. Avanzini
Dip. di Ingegneria dell'Informazione
Università di Padova
{marchet1|avanzini}@dei.unipd.it

SOMMARIO

Questo lavoro presenta una tecnica per la stima del modello a due masse della corda vocale a partire da un dato flusso glottale tempo-variante. Il modello a due masse è specificato da un certo numero di parametri meccanici di basso livello, calcolati in funzione di quattro parametri articolatori (livelli di attivazione di tre muscoli laringali e pressione subglottale). Le forme d'onda del flusso glottale, sintetizzate dal modello, sono caratterizzate da un insieme di parametri acustici per la quantificazione della sorgente vocale. Misurando un flusso glottale di riferimento viene data una sequenza di parametri acustici e, impiegando la programmazione dinamica e l'interpolazione con reti RBF (Radial Basis Function Networks), si derivano i parametri di attivazione muscolare che portano alla risintesi del flusso glottale di partenza.

1. INTRODUZIONE

Un problema di ricerca aperto nella modellazione fisica delle corde vocali a bassa dimensionalità è la relazione tra i parametri dei modelli ed i parametri acustici relativi alla *voice quality*.

Un recente lavoro [1] ha studiato la sensibilità dei parametri acustici del flusso alla variazione dei parametri fisici di un modello a due masse, dando indicazioni sul comportamento del modello per la simulazione delle diverse voice qualities. I parametri del modello (di basso livello: masse, costanti elastiche, ecc.) non sono però controllati in modo volontario dal parlatore: è necessario uno spazio di controllo fisiologicamente motivato per il modello. Una questione affine è il cosiddetto "problema inverso", ovvero il problema di stimare parametri di controllo tempo-varianti da usare come ingresso al modello fisico, così da risintetizzare un segnale acustico *target*. Questo implica l'inversione di un sistema dinamico non lineare con un elevato numero di parametri; la soluzione, inoltre, può non essere univoca. A questo proposito, per evitare la non-univocità, è possibile lavorare su sequenze temporali di *frames* acustici e stimare i parametri articolatori mediante la minimizzazione di una funzione di costo che includa una componente di "sforzo articolatorio". Questo approccio è stato applicato in [2] alla soluzione del problema inverso relativo ad un modello articolatorio del tratto vocale.

Questo articolo presenta una procedura per la stima di un modello a due masse delle corde vocali [3] a partire dai parametri acustici, tempo-varianti, di un flusso glottale target; il modello è specificato da un vasto numero di parametri fisici di basso livello. Questi parametri fisici, calcolati da un livello aggiuntivo di modellazione, vengono ottenuti come funzione di quattro parametri articolatori (tre livelli di attivazioni di muscoli laringali e la pressione subglottale) [4]. Le forme d'onda dei flussi glottali sintetizzati dal modello sono caratterizzate da un insieme di parametri acustici: frequenza fondamentale F_0, *open quotient OQ*, *speed quotient SQ*, *return quotient RQ*, *normalized amplitude quotient NAQ* [5], ecc. Questi vengono riconosciuti in letteratura come una tipica quantificazione della sorgente vocale [6].

Esistono quindi tre distinti spazi di parametri, legati tra loro: articolatorio, fisico ed acustico. Questo lavoro affronta il problema della mappatura dei parametri acustici nelle loro controparti articolatorie. *Frames* temporali del segnale di flusso glottale vengono caratterizzati mediante sequenze di parametri acustici; viene poi sviluppata una metodologia per derivare le corrispondenti sequenze di parametri articolatori usando tecniche di programmazione dinamica. La procedura è ulteriormente migliorata usando le *Radial Basis Function Networks* (RBFN, reti neurali a funzioni base radiali) per interpolare i punti dello spazio articolatorio. I risultati ottenuti mostrano che il modello fisico, controllato con i parametri stimati, è in grado di risintetizzare un segnale di flusso target con buona accuratezza.

La Sezione 2 descrive il modello fisico usato nel presente lavoro mentre la Sez. 3 dettaglia le tecniche usate per stimare il modello a partire da una segnale glottale tempo-variante target. I risultati, le limitazioni e le lacune dell'approccio proposto vengono discusse in Sez. 4

2. IL MODELLO FISICO

L'analisi sviluppata nelle prossime sezioni è basata sul modello a due masse presentato in [3] e raffigurato in Fig. 1. Il modello assume il flusso, dalla regione subglottale al *punto di separazione* z_s nella glottide, come unidimensionale, quasi-stazionario, privo di attriti e incomprimibile; nella glottide avviene la separazione del flusso e la formazione del getto libero. Non è prevista alcuna risalita di pressione all'uscita della glottide. Il punto di separazione z_s si ha quando l'area glottale $a(z)$ supera il suo valore mini-

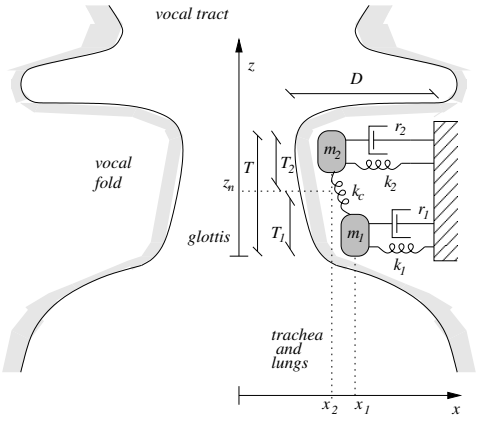


Figura 1. A destra: diagramma della corda vocale, trachea e tratto vocale supraglottale; a sinistra: modello a due masse della corda vocale.

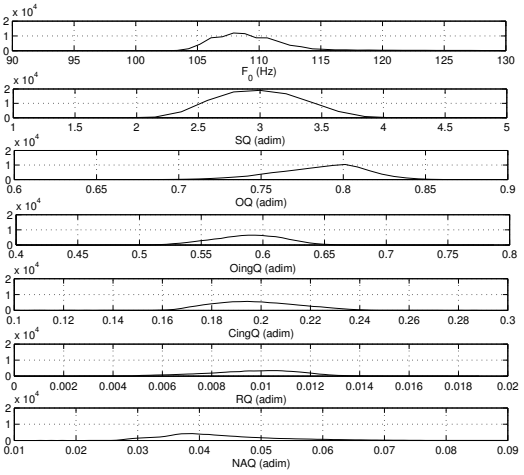


Figura 2. Distribuzione dei parametri acustici, visti singolarmente, nel codebook diretto.

mo di una certa quantità $(10 - 20\%)$ [3]. Introducendo una *costante di separazione* s (nell'intorno $1.1 - 1.2$) la separazione si ha quando l'area glottale assume il valore $a_s = \min(sa_1, a_2)$.

Il tratto vocale è modellato come un carico inertivo. La colonna d'aria in esso presente, assumendo frequenze fondamentali molto più basse del primo formante, agisce approssimativamente come una massa accelerata in modo unitario; la pressione all'ingresso del tratto vocale può essere scritta come $p_v(t) = Ru(t) + I\dot{u}(t)$, dove R, I sono rispettivamente la resistenza e l'inertanza di ingresso. I valori di R, I sono dati da [7]. Essendo un sistema del primo ordine, questo modello non tiene conto delle risonanze del tratto vocale; i suoi effetti più rilevanti sull'oscillazione delle corde vocali vengono comunque descritti con sufficiente accuratezza, in particolare l'abbassamento della soglia di pressione per l'oscillazione [7].

I parametri fisici di basso livello (masse, costanti elastiche, ecc.) non sono controllati dal parlatore: è necessario quindi uno spazio di controllo fisiologicamente motivato, che richiede di stabilire una mappatura tra la fisiologia (attivazioni muscolari) e la fisica (parametri del modello a due masse). Un insieme di regole empiriche, derivate da [8], sono state usate in [4] per il controllo del modello fisico. Le regole legano la geometria delle corde vocali ai livelli di attivazione di tre muscoli: cricotiroidale (a_{CT}), tiroaritenoidale (a_{TA}) e cricoaritenoidale laterale (a_{LC}). Si assume che questi livelli siano normalizzati nell'intorno $[0, 1]$. In questo articolo, inoltre, consideriamo la pressione subglottale p_s. In conclusione il modello fisico è completamente controllato da un insieme di quattro *parametri articolatori*: $a_{CT}, a_{TA}, a_{LC}, p_s$.

3. STIMA DEL MODELLO

3.1 Il *codebook* articolatorio

Il primo passo della procedura di stima consiste nel definire e popolare un *codebook diretto*, in cui ogni vettore di parametri articolatori $[a_{CT}, a_{TA}, a_{LC}, p_s]$ è una "chiave" associata con uno ed uno solo vettore di parametri acustici. A questo scopo è stato condotto un grande numero di simulazioni numeriche del modello a due masse, variandone i parametri articolatori in una griglia densa di valori. Per ogni simulazione i parametri acustici di rilievo del flusso sintetizzato sono stati misurati mediante il toolkit APARAT [9].

Il codebook diretto usato in questo lavoro è stato ottenuto con una griglia in cui a_{CT} e a_{TA} variano nell'intorno $[0, 1]$ con passo fisso 0.05, mentre l'intorno di a_{LC} è $[0.25, 0.5]$ con un passo di 0.025 (la fonazione è ottenibile solo in questa regione). p_s varia nell'intorno $[500, 1500]$ Pa con un passo fissato di 50 Pa. Il codebook risultante contiene 86125 coppie di vettori articolatori/acustici; la Fig. 2 mostra la distribuzione dei 7 parametri acustici nel codebook.

3.2 Inversione del codebook ed accesso dinamico

Per risolvere il problema inverso il codebook diretto è stato invertito, ottenendo il *codebook inverso*. Quest'ultimo però manifesta il problema della non-univocità, ovvero un medesimo vettore acustico può essere associato a uno o *più* vettori articolatori. Affrontiamo il problema lavorando, piuttosto che su singoli vettori, su sequenze temporali di vettori acustici. Queste possono essere ottenute, ad esempio, analizzando un flusso glottale tempo-variante per sezioni (*frame-by-frame*). Data una sequenza di vettori acustici x_k vogliamo ottenere una sequenza "ottima" di vettori articolatori v_k^j nel codebook inverso: come già spiegato, ogni x_k è in principio associato a molti vettori candidati v_k^j a causa del problema di non-univocità. In particolare, eseguiamo una ricerca nello spazio acustico del codebook inverso per trovare i vettori più vicini (distanza euclidea) ai dati x_k; i vettori v_k^j sono quindi i vettori articolatori associati ai vettori acustici del codebook più vicini agli x_k.

La sequenza ottima di parametri articolatori è ottenuta minimizzando una *funzione di costo* formata da tre termini. Un termine *acustico* tiene conto della distanza euclidea tra x_k e la sua versione discretizzata nello spazio acustico del codebook (i vettori trovati mediante ricerca). Un termine *articolatorio* minimizza la distanza euclidea tra v_k^j e v_{k-1}^j, ovvero tra tutte le coppie di vettori articolatori *consecutivi nel tempo*. Questo è il termine chiave della procedura, che permette di ottenere variazioni fluide dei parametri: esso minimizza lo "sforzo articolatorio", in accordo con il comportamento fisiologico dei muscoli. Un termine di *accumulazione* estende il dominio della funzione di costo all'intera sequenza di input, così da garantire l'ottimalità della sequenza articolatoria in senso globale.

La funzione di costo (semplificata) è:

$$f(v_k^j) = \min_{\gamma,\delta}[\tau_1||x_k - c_k^\delta||^2 + \tau_2||v_k^j - v_{k-1}^\gamma||^2 + f(v_{k-1}^\gamma)]$$

dove $\tau_{1,2}$ sono i pesi per i termini acustico ed articolatorio, rispettivamente; c_k^δ sono i vettori acustici discretizzati più vicini agli x_k. Le tecniche di programmazione dinamica sono gli strumenti essenziali per la minimizzazione della funzione di costo: in particolare il termine di accumulazione porterebbe ad una complessità esponenziale, che però è evitata usando questo approccio; la complessità computazionale rimane quindi polinomiale.

3.3 *Clustering* del codebook e interpolazione con RBFN

Un problema della procedura proposta è che i vettori target x_k tipicamente non sono presenti nel codebook inverso, che è discretizzato; ogni v_k^j trovato non è quindi associato con x_k, ma solo con un vettore simile (vicino) ad x_k. Le limitazioni del codebook discreto possono essere superate interpolando lo spazio articolatorio; questo permette di calcolare i vettori articolatori associabili esattamente ad ogni dato x_k.

L'interpolazione usa RBFN (Radial Basis Function Networks, reti a funzione base radiale) [10]. Dal momento che le RBFN interpolano solo funzioni e non possono gestire multimappe, il codebook inverso deve essere manipolato in modo da evitare il problema di non-univocità. Abbiamo sviluppato un innovativo algoritmo che suddivide il codebook in insiemi (*clusters*) acustici ed in sottoinsiemi (*subclusters*) articolatori; ogni cluster è associato ad uno o più subclusters. L'algoritmo garantisce che per tutti i vettori acustici in un dato cluster ci sarà solo un (o nessun) vettore articolatorio in ogni subcluster associato. In questo modo in ogni subcluster il codebook, così suddiviso, dà una mappatura univoca, necessaria per il corretto funzionamento delle RBFN.

L'algoritmo di *clustering* prima suddivide lo spazio acustico in clusters C_i usando una tecnica standard. Vengono generati dei vettori acustici casuali, tanti quanti il numero di clusters desiderato, che seguito vengono variati con una procedura iterativa [11], diventando così i centroidi dei clusters. I centroidi vengono ripetutamente spostati in modo tale che la somma delle distanze tra ogni centroide ed i vettori a lui associati (ovvero associati al cluster del centroide) sia minimizzata. I clusters C_i sono infine costituiti associando ciascun vettore acustico al centroide più vicino. Per ottenere una distribuzione uniforme dei vettori in ogni cluster, la procedura iterativa è applicata in due passi; inoltre, per assicurare un certo grado di sovrapposizione tra i clusters (necessario per il corretto funzionamento dell'ottimizzazione), i vettori più vicini al confine tra due clusters adiacenti vengono replicati in entrambi.

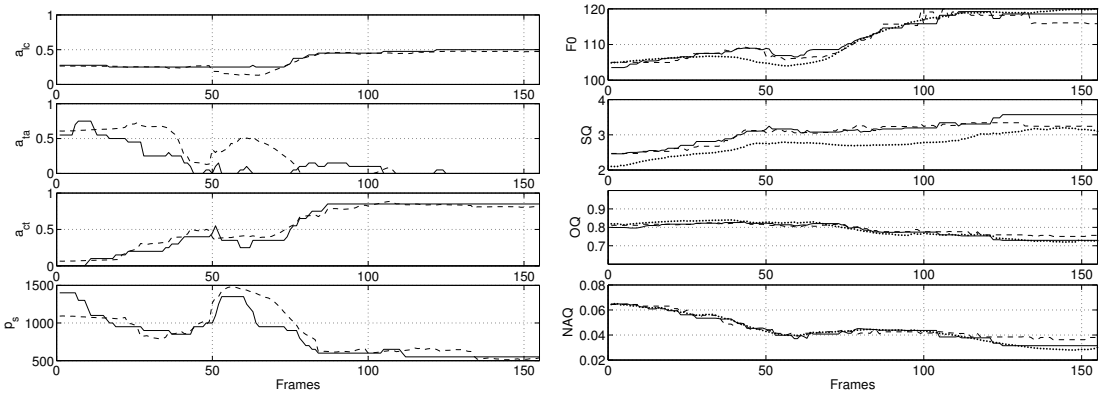
Quando i clusters acustici C_i sono costituiti, l'algoritmo di clustering determina gli s subclusters articolatori S_j^i ($j = 1 \dots s$) associati a ciascun C_i. s è uguale al massimo numero di vettori articolatori associati al medesimo vettore acustico x^* in C_i. Ogni vettore articolatorio associato ad x^* è posto in un subcluster distinto ed usato come "seme". I rimanenti vettori articolatori vengono allocati come segue. Quando diversi vettori articolatori v_k^j sono associati al medesimo vettore acustico x_k, ogni v_k^j è assegnato ad un differente subcluster, scegliendo quello che ha il centroide *articolatorio* più vicino. La posizione di tale centroide è aggiornata in seguito ad ogni aggiunta di nuovi vettori.

Avendo determinato i clusters C_i ed associato ciascuno con uno o più subclusters S_j^i, all'interno di ogni S_j^i costruiamo quattro diverse reti RBFN per interpolare ogni dimensione dello spazio articolatorio. Ogni vettore acustico associato al subcluster è usato come centro per una funzione base della RBFN (nel nostro caso funzioni gaussiane). I valori per i parametri di ciascuna funzione base (deviazione standard, ecc.) sono stati determinati con una estesa serie di sperimentazioni sul codebook. Dopo aver determinato tutte le RBFN è possibile interpolare lo spazio articolatorio. La seguente procedura è utilizzata per passare alla programmazione dinamica i vettori interpolati. Dato un vettore acustico troviamo i k cluster acustici a lui più vicini, e tutti i subclusters ad essi associati. Il vettore acustico è quindi usato come dato di ingresso per l'insieme delle RBFN presenti in ciascun subcluster. Infine, tutti i vettori articolatori ottenuti dall'interpolazione (tanti quanti il numero di subclusters individuati) vengono passati alla procedura di programmazione dinamica, che procede all'ottimizzazione.

4. RISULTATI E DISCUSSIONE

Gli algoritmi proposti sono stati inizialmente testati e tarati con sequenze artificiali di vettori acustici target. Queste sono state usate come ingresso al sistema per ottenere i corrispondenti parametri articolatori. I risultati ottenuti da questi test preliminari hanno dato due indicazioni; per prima cosa, si è verificato che i segnali sintetici ottenuti controllando il modello fisico con i parametri articolatori ottenuti inseguono bene i vettori acustici target. La seconda indicazione è stata che le attivazioni muscolari e la pressione subglottale, ottenute dal sistema, hanno evoluzioni fisiologicamente attendibili, ovvero presentano variazioni fluide nel tempo. Questi risultati iniziali confermano la validità della funzione di costo usata e della interpolazione con RBFN.

Per verificare gli algoritmi proposti su segnali reali abbiamo realizzato una procedura completa di "sintesi mediante analisi" (*synthesis-by-analysis*). Partendo da una voce registrata (una vocale sostenuta con altezza e voice



(a) (b)

Figura 3. Esempio di procedura di sintesi mediante analisi. (a) Sequenze temporali di parametri articolatori ottenute dalla procedura di ottimizzazione (linea solida: senza RBFN; linea tratteggiata: con RBFN). (b) Sequenze temporali di parametri acustici del flusso glottale (linea a punti: sequenze target estratte da una frase registrata; linea solida: risintesi senza RBFN; linea tratteggiata: risintesi con RBFN).

quality variabili) il segnale è stato sottoposto a filtraggio inverso mediante APARAT. Il flusso glottale stimato è stato analizzato *frame-by-frame*, ottenendo una sequenza di vettori acustici misurati. I corrispondenti vettori articolatori, derivati mediante i procedimenti esposti in Sez. 3, vengono usati per guidare il modello fisico; il flusso glottale risintetizzato viene quindi convoluto con il filtro tempo-variante dei formanti del tratto vocale. Il risultato finale è la risintesi del segnale vocale di partenza, in cui l'evoluzione dell'altezza e della voice quality è simile a quella originale.

La Fig. 3 mostra le prestazioni della procedura di sintesi mediante analisi su una fonazione reale (una /e/ sostenuta). I vettori acustici tempo-varianti ottenuti nella risintesi inseguono con buona accuratezza quelli target; test di ascolto informali confermano che la risintesi è qualitativamente simile al segnale originale. In particolare, *NAQ* è solitamente ben inseguito, come visibile in Fig. 3(b). Questo è un risultato positivo poiché *NAQ* è noto essere fortemente correlato alla voice quality [5]. L'effetto dell'impiego delle RBFN può essere notato in Fig. 3(a): le sequenze di vettori articolatori interpolate dalle RBFN hanno variazioni più fluide rispetto alle sequenze ottenute usando la sola programmazione dinamica. Un secondo vantaggio dell'impiego delle RBFN è che il numero di vettori che vengono forniti alla procedura di programmazione dinamica viene significativamente ridotto, portando così ad una riduzione del tempo di calcolo necessario all'ottimizzazione.

I risultati riportati in questo lavoro indicano che l'approccio proposto è efficace nella stima dei parametri di controllo del modello fisico, sia con dati target sintetici, sia con segnali vocali reali; tuttavia, alcune limitazioni affliggono le prestazioni della procedura di stima descritta nell'articolo. Esse sono principalmente legate alle limitazioni intrinseche del modello a due-masse; gli intorni di variazione per i parametri acustici, infatti, sono in genere ridotti (si veda Fig. 2) ed, in alcuni casi, non realistici. *RQ* e *NAQ* in particolare assumono valori eccessivamente bassi; la causa è la limitata capacità del modello di descrivere il flusso con piccole aperture glottali. Questo causa la simulazione di chiusure glottali improvvise e un picco della derivata del flusso eccessivamente alto. La relazione tra i parametri fisici del modello ed i parametri acustici necessita inoltre di essere verificata meglio: ad esempio, la relazione tra p_s ed F_0 osservata nel modello non è in accordo con i risultati riportati in letteratura. Infine, per sfruttare appieno i benefici dell'interpolazione del codebook, è necessario un approccio più sistematico alla determinazione dei parametri delle RBFN.

5. BIBLIOGRAFIA

[1] D. Sciamarella and C. D'Alessandro, "On the acoustic sensitivity of a symmetrical two-mass model of the vocal folds to the variation of control parameters," vol. 90, pp. 746–761, July 2004.

[2] J. Schroeter and M. Sondhi, "Speech coding based on physiological models of speech production," in *Advances in Speech Signal Processing* (S. Furui and M. Sondhi, eds.), pp. 231–263, New York: Dekker, 1992.

[3] N. J. C. Lous, G. C. J. Hofmans, R. N. J. Veldhuis, and A. Hirschberg, "A symmetrical two-mass vocal-fold model coupled to vocal tract and trachea, with application to prosthesis design," vol. 84, pp. 1135–1150, 1998.

[4] F. Avanzini, S. Maratea, and C. Drioli, "Physiological control of low-dimensional glottal models with applications to voice source parameter matching," vol. 92, pp. 731–740, Sept. 2006.

[5] P. Alku, T. Bäckström, and E. Vilkman, "Normalized amplitude quotient for parametrization of the glottal flow," vol. 112, pp. 701–710, Aug. 2002.

[6] P. Alku and E. Vilkman, “A comparison of glottal voice quantification parameters in breathy, normal and pressed phonation of female and male speakers,” vol. 48, pp. 240–254, Sept. 1996.

[7] I. R. Titze and B. H. Story, “Acoustic interactions of the voice source with the lower vocal tract,” vol. 101, pp. 2234–2243, Apr. 1997.

[8] I. R. Titze and B. H. Story, “Rules for controlling low-dimensional vocal fold models with muscle activation,” vol. 112, pp. 1064–1027, Sept. 2002.

[9] M. Airas, H. Pulakka, T. Bäckström, and P. Alku, “A toolkit for voice inverse filtering and parametrisation,” in *Proc. 9th European Conf. on Speech Communication and Technology (Interspeech’2005 - Eurospeech)*, (Lisbon), pp. 2145–2148, Sept. 2005.

[10] T. Poggio and F. Girosi, “Networks for approximation and learning,” *Proceedings of the IEEE*, vol. 78, pp. 1481–1497, Sept. 1990.

[11] A. Gercho and R. M. Gray, *Vector quantization and signal compression*. The Kluwer international series in engineering and computer science, Kluwer, 1992. Boston.