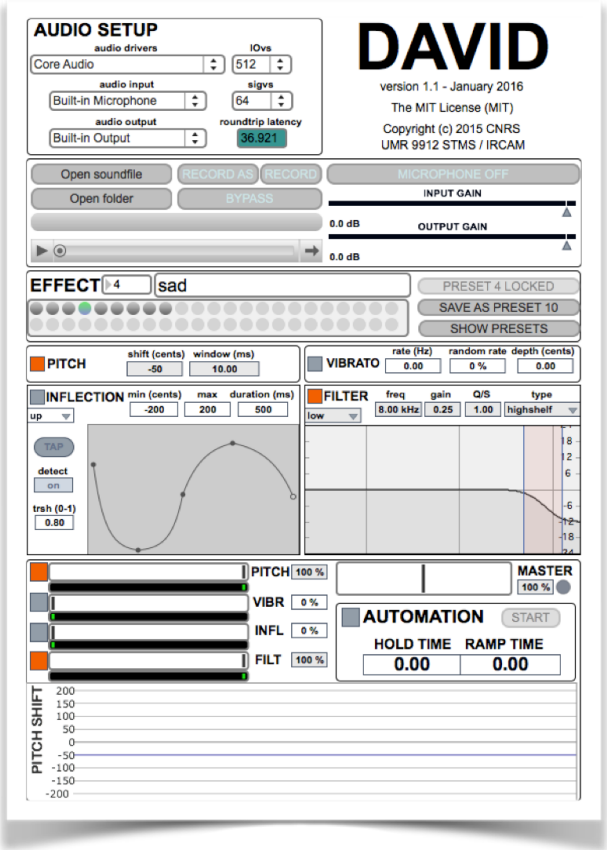
Andrea Gozzi

D.A.V.I.D.
Il software audio che dà voce alle emozioni e viceversa.

Recentemente un gruppo di ricercatori all'interno dell'IRCAM (http://www.ircam.fr/) riunitisi sotto la sigla CREAM (http://cream.ircam.fr/) e autodefinitisi *audio geeks on a mission to the Amygdala* sta indagando il modo in cui la musica influisce sulle emozioni umane, muovendosi in un ambito di ricerca multidisciplinare che unisce l'elaborazione del segnale audio alle neuroscienze cognitive. Una delle prime loro ricerche ha avuto come focus la voce umana.

A settembre 2015 è stato pubblicato online il software D.A.V.I.D. (*Da Amazing Voice Inflection Device*), ideato e realizzato da Marco Liuni (http://scholar.google.fr/citations?user=Tz-18gsAAAAJ&hl=en) assieme all'*équipe* CREAM e chiamato così anche in onore di uno dei suoi primi celebri *tester*, il musicista David Byrne.

Il software, una *patch* di Max (Cyclin'74) *open source,* permette attraverso una serie di strumenti di trattamento sonoro digitale di "aggiungere emozioni" ad una voce neutra o senza particolari inflessioni emotive, funzionando sia in tempo reale (con un ritardo di circa 15 millisecondi) sia su campioni audio pre-registrati. Combinando quattro differenti finestre di controllo principali (*pitch*, *vibrato*, *inflection* – cioè un *local pitch filter* ovvero un inviluppo – e *filter*) e la relativa quantità di effetto desiderato per ognuno degli strumenti, è possibile creare diverse inflessioni emotive e salvare i parametri come *preset* (alcuni sono forniti assieme al software) per creare una libreria personale. Si può anche decidere di passare da uno stato emozionale all'altro attraverso un controllo dedicato, ottenendo un passaggio graduale tra più stati in base ad un tempo da specificare.



Qualche esempio pratico. Per ottenere un'inflessione "felice" si è optato per una manipolazione che vede l'utilizzo del *pitchshit* verso tonalità più alte (*highshift*) unita ad un modifica dinamica ottenuta tramite un compressore audio, per dare l'idea di "prossimità e confidenza", unite infine ad una modifica spettrale attraverso un filtro passa-alto per dare l'idea di eccitazione. In senso

contrario, la manipolazione “triste” prevede invece un *pitchshift* verso tonalità basse (*downshift*) e l’utilizzo di un filtro passa-basso. Per dare un’inflessione “impaurita” si è optato per l’utilizzo del vibrato in combinazione con un cambio repentino di dinamica nel tempo per ottenere un suono “tremolante”. Se ne possono ascoltare alcuni esempi a questo indirizzo: http://cream.ircam.fr/?p=44

Abbiamo contattato uno degli sviluppatori del software, Dr. Marco Liuni, per alcune domande.

Qual è stato il lavoro concettuale e tecnico alla base del processo di "analisi" per procedere poi alla realizzazione del software?

MARCO LIUNI: *Il lavoro svolto per D.A.V.I.D. è molto simile a quel che farebbe un sound designer, piuttosto che quello di un ingegnere del segnale. Non abbiamo definito un modello esaustivo per l’espressione di un’emozione, il nostro approccio è stato di scegliere un modello sulla base di evidenze fisiche semplici: ad esempio, un lieve aumento del pitch o un’enfatizzazione delle frequenze acute associato alla gioia, l’inverso alla tristezza. Sulla base di elementi di questo tipo, abbiamo costruito un insieme di moduli elementari, e determinato “ad orecchio” alcune combinazioni che possono essere associate a emozioni specifiche: dopo di che, abbiamo validato il modello, testando i risultati ottenuti con esperienze percettive su un ampio numero di persone, di cultura e lingua diverse (Francia, Giappone, Inghilterra, Svezia).*

In questa fase di analisi avete preso in esame il comportamento vocale di un numero di persone?

M.L: *L’approccio tipico per questo genere di applicazione in elaborazione del segnale è quello di creare un modello, a partire dall’apprendimento automatico dell’espressione vocale di un certo numero di persone, tipicamente attori: si chiede loro di esprimere diverse emozioni in una stessa frase, in modo che la macchina possa apprendere un comportamento di riferimento. Nel caso delle emozioni, le ricerche di cui eravamo a conoscenza hanno incontrato due problemi fondamentali: in primo luogo, persone diverse esprimono la stessa emozione in modo diverso, e anche una stessa persona non esprimerà necessariamente un’emozione sempre in modo identico; inoltre, una stessa persona può esprimere emozioni diverse attraverso un comportamento molto simile. I modelli appresi, di conseguenza, risultano estremamente ambigui.*
Per aggirare questi problemi, non c’è stato alcun apprendimento automatico nella fase di implementazione del software. Si tratta dunque di un modello euristico, basato su alcune evidenze fisiche.

In base a quali criteri e corrispondenze (es. "voce felice = pitch più alto") avete scelto gli strumenti da utilizzare nel software?

M.L: *I moduli più elementari sono trasposizione costante del pitch e filtraggio, associati in modo intuitivo alto-felice, basso-triste. Inoltre esiste un modulo che applica un vibrato continuo, che può variare da periodico a parzialmente aleatorio, associato alla paura. In ultimo, l’inflessione: si tratta di una modulazione locale del pitch attivata da una rilevazione automatica di transitori d’attacco. Tipicamente, ad ogni inizio di frase o in corrispondenza di un fonema più energico, viene attivata una modulazione, il cui profilo è variabile, di una durata intorno ai 500 millisecondi. Un profilo con una rapida ascesa e una più lenta discesa (come una campana asimmetrica) è associato alla gioia, uno pseudo-sinusoidale alla paura.*

D.A.V.I.D. è stato definito "un *Auto-Tune* per le emozioni" (Brian Resnick, per *Vox.com*), sei d'accordo con questa definizione?

M.L: *In chiave giornalistica l’immagine può essere efficace, perché rende in una riga l’idea che io, probabilmente, non sono ancora riuscito qui a rendere in una pagina… La differenza principale è*

che Auto-Tune è un software commerciale, D.A.V.I.D. è un prototipo realizzato da un ricercatore, questo giusto per calmare troppo facili entusiasmi.
D'altra parte in linea di principio, non è un'immagine sbagliata, perché D.A.V.I.D. nasce proprio per attivare un processo di feedback emozionale, ovvero un meccanismo per cui ascoltando la propria voce modificata in "più gioiosa" o "più triste", anche il proprio stato emozionale si altera di conseguenza. Si tratta di cercare una risposta, nell'ambito dell'apparato uditivo, ad un quesito aperto sulla relazione tra "corpo" e "cervello" riguardo alle emozioni: ad esempio, rido e dunque provo gioia, o provo gioia e dunque rido?
Tecnicamente parlando, D.A.V.I.D. e Auto-Tune hanno alcuni limiti simili: se canto completamente fuori tonalità, Auto Tune introdurrà artefatti nel risultato, ovvero il canto modificato suonerà innaturale (processo con cui si possono anche fare ottimi brani, si vedano i Radiohead…). Allo stesso modo, D.A.V.I.D. non è in grado di cambiare il contenuto emozionale di una voce in modo drastico: è possibile aggiungere un carattere emozionale ad una voce neutra, o rinforzare quello di una voce già espressiva; ma una voce triste non diventerà gioiosa.
Infine, i due software si applicano in differenti contesti, sebbene entrambi gli algoritmi agiscano sul pitch: D.A.V.I.D. deve poter essere riascoltato in cuffia da una persona che parla in un microfono, senza accorgersi che la voce è modificata. Questo impone due vincoli specifici: primo, un'estrema rapidità. Il suono deve tornare alle orecchie in meno di 30 millisecondi, tempo oltre il quale il ritardo diventa percepibile, ed estremamente fastidioso per chi parla. Questo impone l'uso di un algoritmo di pitch shift molto semplice, inadatto a trasposizioni elevate. Secondo, la modifica deve essere leggera, tale da non essere percepita come un agente esterno.
Auto-Tune risponde a esigenze differenti, prima tra tutte quella di dover misurare il pitch originale della voce; inoltre deve eseguire trasposizioni alle volte piuttosto elevate, il che necessita di algoritmi di shift più accurati, che accessoriamente permettano la trasposizione delle formanti. Impossibile, in questo contesto, garantire la rapidità di D.A.V.I.D.

D.A.V.I.D. porta questo nome anche per uno dei primi e più celebri *user* del software, David Byrne: qual è stata la sua reazione alla prova?

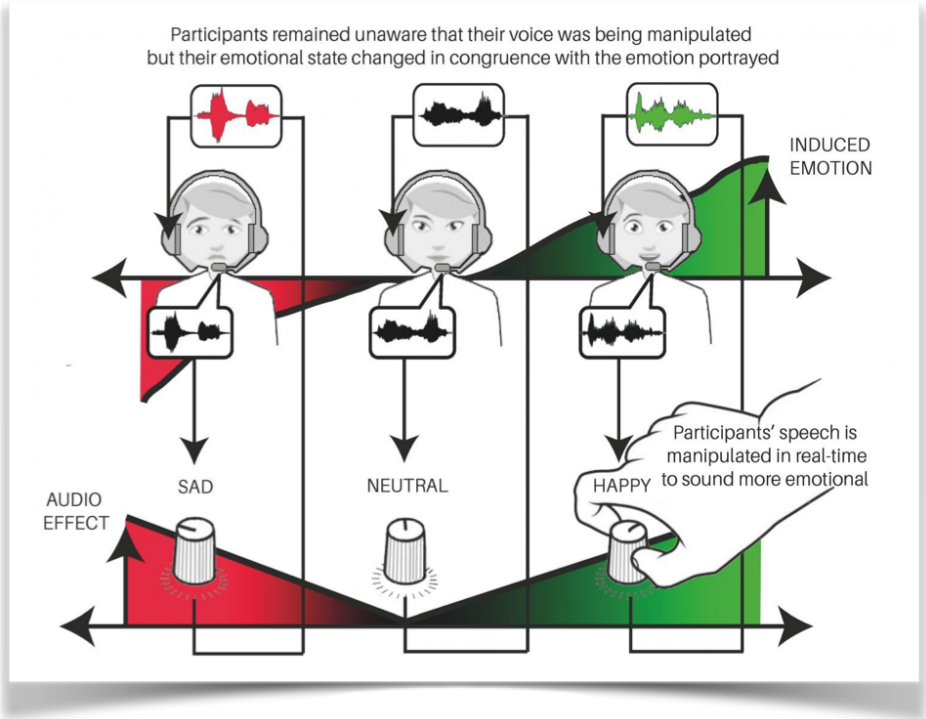
M.L: *In équipe eravamo come bambini alle giostre, e allo stesso tempo avevamo una tensione altissima per il peso di un giudizio di quel calibro… Ho attivato l'effetto con il dito esitante sul mouse, mentre Byrne leggeva un testo, una sensazione simile ai miei primi saggi di chitarra. Dopo pochi secondi, lui si toglie le cuffie, con un sorriso complice, e dice "it's immediate". Davvero una bella soddisfazione, perché si tratta di qualcuno che per professione è portato a riascoltare spesso la propria voce in un contesto simile: se lui è convinto, è un buon inizio.*

Quanto pensi che influirà nei prossimi anni la ricerca sulla trasmissione di "emozioni virtuali" tramite la telecomunicazione?

M.L: *Enormemente. La nostra ricerca parte da uno studio sulla percezione delle proprie emozioni, ed è orientata soprattutto a casi clinici (stati depressivi, ad esempio) in cui percepire se stessi con un'alterazione emozionale può innescare una variazione del reale stato emotivo. Le potenziali applicazioni in musica sono vaste, se si accetta l'ipotesi di poter tracciare paralleli tra la percezione delle emozioni nella voce parlata e nel canto, o nel suono di uno strumento.*
Ma si può anche immediatamente pensare a rendere la voce di un operatore di un call center più gioiosa. E più in generale, la ricerca nell'ambito dei sistemi interattivi capaci di rilevare lo stato emotivo della persona, modificando il proprio comportamento di conseguenza, è molto prolifica. Sono già allo studio robot dotati di "sensibilità", capaci di sostenere un colloquio adattandosi alle reazioni dell'interlocutore.

Come accennava Liuni, questo software è stato poi impiegato in alcuni esperimenti che presentavano come focus principale: la voce, le emozioni comunicate, gli effetti che possono produrre e la consapevolezza dell'influenza che queste hanno sul comportamento umano. Ai partecipanti venivano fatte leggere alcune brevi storie delle scrittore giapponese Haruki Murakami davanti ad un microfono, il segnale audio era poi diffuso tramite cuffie stereo fornitegli in precedenza. Dopo alcuni minuti, senza avvertire il partecipante, si cominciavano ad effettuare

alcune elaborazioni dei segnale sulla sua voce. I risultati sono molto interessanti: quasi tutti i partecipanti modificavano involontariamente la propria voce a seguito del trattamento audio, conformemente al segno del carattere dell'elaborazione (voce triste = stato d'animo triste) e una ridottissima percentuale si è accorta che la propria voce stava venendo processata in qualche modo. Gli esperimenti sono stati svolti presso lo stesso IRCAM, la University College of London in Inghilterra, la Lund University in Svezia e la Waseda University in Giappone, prendendo in esame persone di entrambi i sessi e provenienti da differenti paesi. Seguendo lo stesso principio i risultati si sono rivelati molto simili tra loro, a sostegno che questi processi emotivi e cognitivi siano da considerarsi in maniera transculturale.
Per approfondire: http://www.pnas.org/content/early/2016/01/05/1506552113.full.pdf



Come funziona esattamente il meccanismo di stimolo e risposta in merito alle emozioni vocali è ancora un campo tutto da indagare. Se gli psicologi non hanno ancora una chiara risposta riguardo al processo emozionale e cognitivo legato alla voce, è comunque parere condiviso che le emozioni possano cambiare il suono della nostra voce e che l'emozione trasmessa dal suono della voce di un'altra persona possa influenzare i nostri comportamenti e produrre in noi reazioni emotive: semplificando, una voce "arrabbiata" ad esempio produce ansia, paura o tristezza, mentre una voce "felice" può coinvolgerci e spingerci in quella direzione, esattamente come accade per alcuni comportamenti del corpo e per la prossemica. È certo necessario fare le dovute distinzioni tra emozioni "più ampie" - come la felicità, che può tradursi con diverse sfaccettature di un medesimo effetto - ed emozioni più "direzionali" che, come la paura, hanno un effetto molto connotato.
Essendo questi meccanismi in gran parte inconsci, uno studio in questa direzione può contribuire a sviluppare una maggiore consapevolezza delle emozioni, e aiutarci a capire meglio gli altri e noi stessi. Una ricerca indirizzata su questi temi può anche aiutare in campo medico nella comprensione degli errori di decodifica emotiva. Si pensi ad esempio a possibili finalità terapeutiche nell'approccio ai disturbi dell'umore e ai problemi comportamentali da essi derivati.
Un'indagine con tali strumenti oltre all'*analisi* di un determinato comportamento potrebbe portare anche a un'ipotesi di produzione sintetica (ma non di *sintesi* nella sua accezione classica utilizzata in musica elettronica) dello stesso o di alcuni meccanismi "di innesco". Si provi ad esempio a pensare alla possibilità di donare delle emozioni trasmettendole tramite comunicazione virtuale o associandole utilizzando risorse audio alla telescrittura. Un campo di indagine che prende sempre più piede e ancora decisamente tutto da esplorare.