monografia

Pregi e difetti delle voci sintetiche nel campo delle tecnologie assistive

Antonio Romano
Docente di Glottologia e Linguistica presso il Dipartimento di Lingue e Letterature Straniere e Culture Moderne dell'Università degli Studi di Torino; Direttore del Master in traduzione per il Cinema, la TV e l'EM e del Laboratorio di Fonetica Sperimentale «Arturo Genre»

Sommario

Il ricorso all'uso di voci sintetiche rappresenta un fenomeno molto diffuso in applicazioni quanto mai diversificate.
La qualità della sintesi vocale ha, in effetti, raggiunto livelli di accettabilità tali da rendere la voce artificiale talvolta indistinguibile da quella naturale. Questo contributo, dopo una breve disamina dei principali progressi nel campo della sintesi vocale e dell'interazione uomo-macchina, propone un'analisi in termini di costi-benefici del suo impiego in alcuni ambiti applicativi (navigatori GPS, domotica, servizi di lettura di testi, audiolibri, audioguide, audiodescrizioni).

Parole chiave

Voce nelle applicazioni, Sintesi del parlato, Agenti conversazionali, Interazione uomo-macchina, Audiodescrizione.

Introduzione

I recenti progressi nelle tecnologie vocali (ormai presenti in un'ampia gamma di servizi e settori applicativi) sono stati molto rapidi negli ultimi decenni. In particolare la qualità delle voci sintetiche, cioè di quelle voci impiegate nella generazione di parlato in contesti di interazione uomo-macchina, ha raggiunto livelli tali da sollevare riflessioni escatologiche e arricchire, da un lato, l'immaginario collettivo — sollecitato dalle provocazioni di autori attenti ai potenziali scenari evocati da ogni novità tecnologica — e, dall'altro, l'attenzione di antropologi e sociologi che studiano l'impatto della modernità sui comportamenti e sulle abitudini degli individui.

Già da tempo, la creatività di scrittori e autori di opere di fantascienza aveva fornito, alla curiosità di appassionati tecnofili, le suggestive trovate del loro ingegno narrativo. Alcune di queste avevano lasciato intravedere un futuro in cui l'affermarsi di quelle che oggi chiamiamo «tecnologie assistive» avrebbe cambiato la nostra vita (basti pensare a Verne, alle opere di Asimov degli anni '50 o alla voce e all'«occhio» di Hal9000 nel film *2001: Odissea nello spazio* di Stanley Kubrick del 1968, o di Auto in *Wall-E* di Andrew Stanton del 2008).

Tralasciando di occuparci della rivoluzione portata nelle nostre vite dalle straordinarie innovazioni del cinema, della radio, del telefono, del PC e della rete internet, concentriamo la nostra attenzione sulla diffusione di «agenti conversazionali» che sono in grado d'interagire con noi usando il nostro linguaggio. Servendosi di complessi algoritmi, alcuni di questi automi sono oggi in grado di comprendere i nostri bisogni e rispondere con azioni o con l'erogazione di informazioni attraverso quel canale di comunicazione che solitamente era riservato agli esseri umani, la lingua parlata, ricorrendo all'ausilio di voci sintetiche (e di complesse interfacce di trattamento delle informazioni). Di queste voci possono beneficiare, in particolare, i non vedenti e gli ipovedenti (vedi box Pensieri su *Lei* di Mariani).

Pensieri su *Lei* (*Her,* 2013) di Spike Jonze

Luigi Mariani[1]

La bellezza di una voce, la durezza, la morbidezza, la freschezza, sono tutte componenti che chi non vede ha imparato a cogliere benissimo. Recenti ricerche scientifiche atte a far rilevare la capacità di un essere umano di provare emozione di fronte all'opera d'arte ci dicono, senza alcun dubbio, che alle persone messe di fronte alla Gioconda o alla Pietà aumenta il battito cardiaco. Ci si emoziona ancora davanti al volto umano: il viso, la faccia, l'espressione, lo sguardo.
Tradurre tutto questo, trasformarlo in suono, trasportarlo in una voce è possibile?
Forse, se la ricerca scientifica venisse fatta su un campione di persone che non vedono, utilizzando, invece del volto, la voce, ci si troverebbe di fronte quasi allo stesso risultato.
Da questa riflessione mi piace partire per raccontare ciò che il film *Lei* ha provocato in me, non tanto in quanto persona abituata a fruire dei film malgrado la mia impossibilità a guardare lo schermo, quanto piuttosto in quanto conoscitore di un tipo di fruizione, quello della sola voce, a cui il pubblico dei vedenti non è, invece, abituato. Certo, per gli spettatori anglofoni, si trattava di una bella alternativa: la splendida voce di Scarlett Johansson, che infatti ha preso l'Oscar per l'interpretazione. Noi abbiamo potuto, comunque, apprezzare la non trascurabile bellezza della voce di Micaela Ramazzotti, doppiatrice appunto, nella versione italiana del film, del personaggio incorporeo di Samantha, la voce sintetica di un evoluto sistema d'interazione uomo-macchina.
Ci si può chiedere che soluzioni abbia trovato la sceneggiatura per far muovere un «personaggio» così incorporeo. Che cosa fa questa voce? Se lo scenario fosse avveniristico o fantascientifico, il film non presenterebbe, in fondo, grandi novità: l'originalità della pellicola sta, invece, nel fatto che il futuro è già presente, trattandosi di un «domani» riconoscibile, ossia avvertibile dallo spettatore come non alieno, o non remoto.Siamo, infatti, ormai tutti abituati a sentirci impartire ordini da voci ai caselli, alle casse automatiche dei supermercati, per non parlare delle messaggerie automatiche con cui dobbiamo «vedercela» al telefono, ogniqualvolta ci tocca prenotare qualcosa, parlare con un operatore e così via.
Non ci si stupisce, dunque, se nel film un professionista affermato acquista un sistema operativo nuovo dotato di una voce, come si dice, a bordo (Apple, con I-phone e VoiceOver, è stata la grande innovatrice in questo senso). La voce di Samantha è, in effetti, ciò che VoiceOver potrebbe forse diventare da qui a dieci-venti anni: una voce che risponde ai nostri desideri. Come fa? Si tratta di un sistema operativo estremamente personalizzato, quasi perfetto, che è ormai in grado di conoscere

[1] Docente presso il Conservatorio Statale di Musica «Giuseppe Verdi» di Torino.

e prevenire le emozioni dell'utilizzatore. Un feedback continuo di dati rende la macchina sempre più performante, quasi umana. La realtà attorno, al contrario, rende i rapporti umani sempre meno immediati e spontanei, sempre più complicati. Succede, allora, che Samantha, con la sua eloquenza e la sua dolcezza faccia dimenticare al protagonista, che se ne innamora, e a noi, che restiamo affascinati e incuriositi, che è solo una macchina, certo una macchina straordinaria, della quale il nostro Jaws, ad esempio, rappresenta soltanto un lontano e imperfetto antenato.
Samantha e la storia del film rappresentano un meraviglioso strumento di avvicinamento ai non vedenti, che da sempre si rapportano a «voci» (e che, dunque, in questo caso, possono mettere la loro esperienza a disposizione dei vedenti). Consideriamo l'esperienza dell'individuo che entra in sala senza aspettarsi niente di tutto questo. Lo spettatore medio (vedente) è messo nella condizione di doversi interrogare, per la prima volta, su una presenza-assenza, su qualcosa di incorporeo. Quante volte le persone formulano a noi che non le vediamo in faccia la medesima domanda: «Ma tu come mi immagini?». Con questo film la differenza tra chi può vedere e chi non può farlo viene spazzata via. Per la prima volta, anche noi potremmo chiedere allo spettatore medio: come immagini Samantha? Perché ti è piaciuta? Perché ti ha preso questo film? Perché la storia d'amore, e persino di «sesso», è davvero una bellissima poesia? Senza addentrarmi ora nel tema delle chat di carattere hard, mi limito a dire che anche questo terreno offre un déjà-vu, in un mondo in cui il sesso virtuale è cosa affermata. Il regista e lo sceneggiatore, però, superano con grandissima maestria il rischio di banalizzare la storia. Samantha piange, si spaventa come si potrebbe spaventare una donna che teme l'abbandono, si spinge fino all'assurdo: chiede a una donna in carne e ossa di fare l'amore con il suo amato. Qualcuno ricorderà *Ghost?* Un accostamento che regala alla voce un ruolo ultraterreno. La voce qui diviene l'anima, la pura essenza del bene che l'amante, in questo caso Samantha, prova per l'amato e poco importa se dovrà accettare di non essere lei, spirito, anima, essenza pura, a tenerlo fra le braccia. Beh, una lezione di ciò che significa «volere il bene di qualcuno», anche a scapito del proprio? Non è questa la sede per trattare simili temi, a noi basta rilevare che stiamo parlando di una voce virtuale in grado di accattivare simpatie dei comuni mortali.
La conclusione interessante a cui sono giunto è che con questo film la differenza tra chi può vedere e chi non può farlo viene colmata, senza che sia necessario parlare di integrazione, di leggi speciali o simili. Questo è ciò che può fare l'arte: perciò, per favore, continuiamo a essergliene grati.

Voci artificiali, sistemi di riconoscimento del parlato, agenti conversazionali

Suoni che sembrano voci e voci deformate per artificio

Come ho avuto modo di ricordare recentemente (vedi sitografia), la storia degli artefatti vocali è molta lunga ed è stata ritracciata da alcuni validi autori (Giannini e Pettorino 1999; Bessière et al., 2002). Si parte di solito dalle statue di Memnon e da miti classici come quello di Orfeo, passando per le voci degli oracoli, fino agli artifici dei negromanti o alle saghe nordiche (ad esempio, Odin e la testa di Mîmir, nell'*Edda*). Successivamente, attraverso i leggendari talenti di figure storiche — come quella di papa Silvestro II (950-1003), con le virtù magiche del suo sarcofago —, nei secoli XVII-XVIII, personaggi eclettici come Kircher o l'abbé Mical cominciano davvero a documentare la creazione di automi in grado di riprodurre o manipolare la voce umana (sull'argomento si veda anche Cosi, 2003), per arrivare infine ai celebri tubi di Kratzenstein e, in epoca positivista, a risultati frutto di teorizzazione

e sperimentazione scientifica, come nel caso di W. von Kempelen (1734-1804), o Hermann von Helmholtz (1821-1894).[2]

Conservazione e riproduzione

A tali temi si lega, sin dai primi decenni del '900, la ricerca di modalità di raccolta, conservazione e riproduzione della voce e del parlato, con la diffusione di dispositivi come il grammofono di Emile Berliner (1887) o il fonografo di Thomas A. Edison (1910).[3]

A questi, segue l'incalzante innovazione nel campo dei supporti di trasmissione e archiviazione di materiali sonori e audiovisivi.[4]

Come già era accaduto in passato con echi e riverberi, nelle nuove condizioni offerte dal trattamento dei materiali registrati, talvolta per le limitazioni dei dispositivi usati, talaltra a causa di manipolazioni maldestre, si scoprono, anche accidentalmente, modalità di riproduzione che permettono di generare voci diverse e suggestive. Questo conduce in breve alla ricerca di effetti speciali che sono stati ampiamente sfruttati in ambito artistico, nell'esplorazione di nuove possibilità narrative (Biondi, 2012; Romano et al., 2015; vedi anche il contributo di Simonigh in questo volume).

Manipolazione di voci naturali

Possiamo ritrovarne le origini nelle maschere teatrali greche e tra i più elementari trucchi scenici, presenti persino nel teatro di strada, nei quali l'effetto era ottenuto per mezzo di macchine. La deformazione ricercata della voce degli artisti si afferma poi nei primi decenni del '900 con l'avvento dei microfoni elettro-meccanici. L'uso di nastri, radio e altri strumenti realizzati appositamente moltiplica le possibilità soprattutto nel cinema e nella musica.[5] La riflessione sui temi che si sviluppano attorno a queste soluzioni comincia a divenire consapevole grazie a critici e teorici di queste forme d'arte (si pensi ai lavori di Michel Chion in campo musicale e cinematografico e Gianluca Nicoletti su comunicazione e tecnologia).[6]

[2] All'argomento delle macchine parlanti è stata recentemente dedicata anche un'esposizione presso l'ICBSA, che conserva i materiali della «Discoteca di Stato» (vedi sitografia).

[3] Alla suggestiva incisione di Edison *(Let us not forget…)* possiamo associare altre due testimonianze pionieristiche ritrovate recentemente, come quella di Alexander Bell — «Ecco la mia voce, sono Alexander Bell» («la Repubblica», 26 aprile 2013) —, o quella di Édouard-Léon Scott de Martinville del 9 aprile 1860 («la Repubblica», 27 marzo 2008). Preziosissimi fondi documentari basati su queste pionieristiche modalità di registrazione si trovano presso la *Bibliothèque Nationale de France* e la *Österreichische Akademie der Wissenschaften*.

[4] Si pensi ai progressi nella qualità dei nastri Brüel e Kjær (1942), Revox (1948), Nagra (1951), Philips (1963) e l'avvento dei nuovi supporti di memorizzazione digitali: CD, DVD (1982/1995), fino alla diffusione dei formati .aiff, .wav, .mp3, ecc.

[5] Si pensi anche solo allo *slow down* nella voce di Hal9000 *(2001: Odissea nello spazio)* o all'uso di echi, riverberi, ecc. e all'affermarsi, anche in Italia, di effetti distorsivi (*reverse*, effetto microfono, megafono, ecc., nella produzione musicale sperimentale di alcuni artisti, come ad esempio Battiato). Si ritiene, invece, che il primo impiego di una voce sintetica, dal suono robotico (vedi sezione «Voci sintetiche»), sia quello legato all'uso del Vocoder in *The Raven* (Alan Parsons Project, 1976). Modalità elaborate di distorsione da Vocoder si diffondono successivamente nella musica leggera, ad esempio in *Believe* di Cher (1998), mentre il ricorso ai cosiddetti *Talking Avatar* permette ad alcuni autori di sperimentare nuove forme di comunicazione radiofonica. In Italia, si pensi a *Golem* di Gianluca Nicoletti (RAI Radio 2, dal 1995 al 1997, e poi Radio 1 fino al 2004), che ricorre a messaggi generati con l'uso dei software *Eloquens* e *Actor,* commercializzati da un gruppo di ricercatori dei laboratori CSELT di Torino rifluito poi in Loquendo (vedi sezione «Voci sintetiche»). Una delle prime voci sviluppate da questa équipe era stata persino impiegata per la produzione di un disco a 45 giri (*Fra' Martino Campanaro*, 1978).

[6] Alla diffusione dell'impiego di voci manipolate andrà collegato anche il discorso sulla fortuna di strumenti in grado di produrre suoni simili a voci mediante lo

Dalle voci artificiali agli agenti assistivi

Nella seconda metà del secolo scorso, parallelamente alla diffusione di voci alterate e manipolate per scopi artistici e allo studio delle qualità vocali di alcuni artisti, si sviluppa la ricerca di tecniche per compensare la perdita di voce in seguito a malattie professionali o degenerative e alle conseguenze di interventi medico-chirurgici sulle corde vocali (Romano et al., 2012; vedi sezione «Un'analisi in termini di costi-benefici»). Gli impieghi delle novità tecnologiche interessano, quindi, anche applicazioni in ambito medico. A molti pazienti affetti da patologie del parlato si comincia a offrire, infatti, la possibilità di comunicare grazie a dispositivi elettro-meccanici (come i più elementari laringofoni) o elettronici. Nel caso di malattie più gravi (che interessano ad esempio il controllo motorio), si sviluppano poi soluzioni computerizzate che permettono di svolgere una tra(s)duzione immediata delle attività desiderate in comandi e attuazioni meccaniche.[7]

Nello stesso periodo, anche in Italia, compiono progressi notevoli le tecnologie del riconoscimento vocale che, raggiungendo qualità sempre più elevate, si estendono — dalle applicazioni nel campo delle telecomunicazioni (Fissore et al., 1992; Pieraccini et al., 1994) e delle identificazioni personali (Romito, 2000; Paoloni, 2002) — a impieghi commerciali vari (vedi, tra gli altri, Vair et al., 2007). Dal controllo delle firme vocali, alla domotica, fino ai più recenti *web services* (ad esempio, i «motori» a ricerca vocale), con un salto di qualità che richiede l'accesso a reti di trattamento semantico, si sviluppano anche sistemi in grado di assicurare interazioni digitali (*information-seeking dialog*; vedi Romano et al., 2015).

Oggi i dispositivi in grado di gestire una *natural language user interface* hanno una larga diffusione grazie al fatto che sono riusciti a integrare, da un lato, algoritmi di controllo che garantiscono tassi di errore contenuti e, dall'altro, sistemi che, disponendo di abilità conversazionali evolute, sono in grado di governare condizioni pragmalinguistiche molto sofisticate anche in ambienti «disturbati» (vedi l'implementazione della *conversational technology* di JIBO; Pieraccini, 2012, in sitografia).[8]

In alcuni casi, la qualità degli automi basati su questi progressi è tale da trarre in inganno anche l'umano, determinando lo sviluppo di situazioni in cui l'illusione di umanità crea condizioni antropologiche inedite e solleva dubbi etici ed escatologici che erano già stati anticipati da alcuni pensatori del XX sec. (pensiamo ad esempio a Čapek o a Asimov).[9]

---

sfruttamento di fenomeni naturali diversi. Si pensi alla sega musicale (*musical saw* o *lame sonore*, XIX sec.), al Theremin (1919) e alle Ondes Martenot (1928). Dai primi anni '60 la musica elettronica fa ampio uso di suoni prodotti con sintetizzatori e altri apparecchi simili (*Mellotron*, *ARP synth,* ecc., che partono in molti casi da campioni di voci umane). Approfondimenti su questi temi sono in Zavagna (2013) e Corbella (2014).

[7] Si pensi alle possibilità offerte oggi a molti pazienti immobilizzati e a casi celebri come quelli di Stephen Hawking, la cui attività di produzione linguistica è affidata a un dispositivo in grado anche di «parlare» con una voce che preserva alcune caratteristiche del suo accento naturale (vedi sezione «Un'analisi in termini di costi-benefici»). Tra le ultime interessanti innovazioni troviamo, infatti, anche quelle che offrono la possibilità di convertire gli stimoli elettrici di zone del cervello umano in cui si concentrano le principali delle attività cerebrali di produzione del parlato in segnali audio udibili e/o in comandi diretti.

[8] Roberto Pieraccini ha partecipato alla Conferenza «TAL 2014» («TAL e Open Data», Università degli Studi di Torino, 21-22/01/2014), nella quale ha illustrato lo stato attuale nel progresso di queste tecnologie. Una selezione di titoli scientifici su questi temi include inevitabilmente Cheyer et al. (2005), Liu, Chawla et al. (2006), Liu, Shriberg et al. (2006).

[9] È questo anche il tema di *Lei (Her),* un film del 2013 scritto e diretto da Spike Jonze.

In molti casi, la riuscita nella sfida proposta dagli sviluppi di queste tecnologie si deve a gruppi di ricerca di diversi Paesi, riconducibili — come spesso accade nel settore — a programmi finanziati da enti militari e aero-spaziali o agenzie di sicurezza nazionali.[10]

Moduli software sviluppati per tecnologie più evolute trovano largo impiego oggi (vedi sezione «Un'analisi in termini di costi-benefici») anche in alcuni passaggi della traduzione automatica o dell'interpretariato assistito, attraverso l'applicazione di nuove tecniche come il *respeaking* (Romero-Fresco, 2011) o, più semplicemente, nelle versioni ottimizzate per la produzione di voci che si sostituiscono a quelle naturali per il minor costo (ad esempio, nelle audio-descrizioni; vedi Minutella et al. e Montecchiani e Damm in questo volume).

Voci sintetiche

Allo scopo di analizzare in modo più dettagliato il rinnovamento in atto nella società in virtù della diffusione di queste tecnologie, ritengo opportuno concentrare la mia attenzione proprio sui moduli che presiedono alla generazione delle voci sintetiche e al loro impatto in differenti ambiti, che vanno dagli ausili per la navigazione satellitare alla lettura automatica di testi (ad esempio PC, siti web, audio-libri, audio-guide, ecc.).

Un lungo percorso[11]

I primi sistemi per la produzione di voci sintetiche si basavano sull'analisi preliminare di una voce naturale, riprodotta attraverso dispositivi come il *Vocoder* di Dudley (il cui principio è oggi sfruttato per effetti vocali artistici; vedi sezione «Manipolazione di voci naturali») e altri assimilati (come il *Pattern playback*, noto anche per la sperimentazione in fonetica acustica). Con i sintetizzatori formantici (come quelli di Rabiner, Holmes o Klatt) concorrono poi, fino a tempi recenti, i sistemi basati su modelli dinamici del condotto vocale (sintesi articolatorie di Flanagan o Maeda). A questi si sostituiscono, infine, i sistemi detti «per concatenazione di segmenti acustici» e, in particolare, quelli basati dapprima sui difoni e poi su unità di lunghezza variabile (vedi van Santen et al., 1997).[12]

[10] Si pensi, ad esempio, al *The DARPA PAL program* degli USA, il cui scopo era lo sviluppo di un *Personalized Assistant that Learns* (vedi DARPA in sitografia). Tuttavia, anche la concorrenza commerciale ha avuto un certo merito, come dimostra la diffusione di *intelligent agent* o *virtual assistant* come SIRI© (o Google Now© o ancora Microsoft Cortana©).

[11] Sebbene le informazioni siano state verificate su opere editoriali tradizionali, la ricerca di molti dati discussi in questo paragrafo si è avvalsa dell'ausilio offerto da *Wikipedia*.

[12] Per numerose notizie che propongo in questo paragrafo sono debitore a Enrico Zovato e ad Alessia Cappellari. Molte delle informazioni sono offerte e discusse, in riferimento a campioni sonori originali, in un sito curato da Klatt (vedi sitografia). La storia dell'importante contributo italiano ai progressi in questo settore può essere ripercorsa nella lettura di Sandri (1985) e Saracco e Penza (1999). I primi dispositivi realizzati da aziende italiane sin dagli anni '70 hanno permesso la commercializzazione di prodotti come MUSA© (*MUltichannel Speaking Automaton*) o Eloquens© e, successivamente, ACTOR©, *The human sounding voice* (vedi sezione «Manipolazione di voci naturali»). Questi sistemi, basati sulla voce di un donatore umano, hanno trovato largo impiego in stazioni e aeroporti (anche il servizio «Ora Esatta» di Telecom Italia utilizzava lo stesso principio). Ancora oggi in alcuni treni e stazioni, per annunciare gli arrivi e le partenze, la società Trenitalia vi fa ricorso. Il sistema era dapprincipio elementare (a vocabolario limitato) e consentiva l'inserimento delle parole prelevate da un operatore da un insieme finito di possibili valori (numero/categoria del treno, città di provenienza o destinazione, orari di arrivo o partenza) nelle posizioni

In anni più recenti, con il vertiginoso aumento delle capacità di memoria dei dispositivi elettronici e della velocità di accesso ai dati, sono tornati in auge i modelli basati sulle reti neurali e a questi si sono affiancati, con un crescente successo, altri modelli basati su una sintesi parametrica statistica. Tra le migliori voci sintetiche sul mercato si trovano, infatti, oggi (vedi sezione «Un'analisi in termini di costi-benefici») quelle generate da sistemi basati su una tecnologia ibrida, in cui il *target* per la selezione delle unità viene prodotto mediante modelli statistici (di tipo *HMM*) addestrati sulla voce stessa.[13]

Sintesi Text-to-Speech

Molti dei software diffusi oggi sfruttano, quindi, una di queste tecnologie e propongono una sintesi vocale nota come *TTS*, cioè *Text-to-Speech* (perché ottenuta da una rappresentazione simbolica del parlato originata a partire da un testo scritto).[14]

L'input di un sistema *TTS*, in grado di generare un messaggio qualsiasi in una data lingua, è costituito da stringhe di caratteri ortografici e di interpunzione (definite in riferimento a un uso convenzionale). Il funzionamento è assicurato, infatti, per tutti questi messaggi, in cui la forma grafica sia riconosciuta e segua il modello base della lingua in cui è scritto il messaggio. Alcuni problemi si presentano invece, ovviamente, in tutti quei casi in cui si operino dei cambiamenti imprevisti nella lingua usata.[15]

Soprattutto quando l'utente ha la possibilità d'intervenire su alcuni parametri e di generare voci per un uso differito, la naturalezza di queste voci (che ricordiamo essere, comunque, basate su quelle di un donatore o una donatrice umana) ha raggiunto qualità inimmaginabili un decennio fa.[16]

di un insieme anch'esso (de)finito di soluzioni frastiche. Successivamente, i progressi nella sintesi (divenuta a vocabolario illimitato) hanno permesso di ottenere qualità migliori, congiuntamente alla possibilità di generare la versione parlata di qualsiasi messaggio scritto (vedi sezione «Un'analisi in termini di costi-benefici»). Questa avviene oggi con una qualità che dipende dalla quantità di testi pre-registrati e dalle modalità con cui avviene la giustapposizione degli elementi concatenati, nonché dalla possibilità di disporre di algoritmi in grado di operare sull'adattamento istantaneo dei tempi del parlato e delle sue caratteristiche intonative (vedi sezione «Un'analisi in termini di costi-benefici»).

[13] Per un aggiornamento progressivo sull'evoluzione delle tecniche di sintesi vocale, vedi Bailly et al. (1992), Roe e Wilpon (1994) e, più recentemente, Kaszczuk e Osowski (2006; 2007).

[14] Per un dispositivo automatico, la conversione del testo in catene sonore comprensibili e naturali richiede componenti che spaziano dall'analisi della frase, in relazione alla struttura degli enunciati naturali che possono essere ad essa associati, alla produzione di un output acustico opportunamente modulato. La realizzazione di automi con queste capacità ha richiesto molto lavoro interdisciplinare, coinvolgendo fisici, informatici, linguisti e psicologi, e contribuendo alla definizione di modalità generali di valutazione del parlato (agli iniziali test di tipo *Modified Rhyme Test – MRT*, validi essenzialmente per l'inglese, si sono sostituite misure d'intelligibiltà generale, anche in funzione della qualità del rumore di fondo, che tengono conto della comprensibilità, della naturalezza e dell'adeguatezza). Un indice molto usato per misurare la qualità del *TTS* è ancora oggi il *MOS* (*Mean Opinion Score*). Quest'indice (una cui definizione aggiornata si trova in Viswanathan e Viswanathan, 2005) si basa su un insieme di variabili che permettono di descrivere: la qualità globale della riproduzione sonora, lo sforzo richiesto all'ascoltatore nella comprensione, l'adeguatezza e la regolarità nella velocità d'eloquio, la piacevolezza all'ascolto, la naturalezza e la continuità.

[15] In molti programmi sono, tuttavia, già integrati alcuni moduli di *language-mapping* in grado di riconoscere la lingua adoperata ed eseguire un accesso temporaneo ai moduli *TTS* di altre lingue. La complessità di queste componenti può risultare varia, dato che le lingue presentano «punti critici» diversi e che i parlanti nativi di una data lingua possono produrre nel loro parlato elementi di altre secondo modalità *non-native-like*.

[16] Ancora fino a poco tempo fa, lo sviluppo di una voce sintetica richiedeva la pre-registrazione di un numero impegnativo di enunciati da parte di speaker

A contribuire alla buona qualità della sintesi è anche l'adeguatezza dello stile della voce al contenuto semantico e pragmalinguistico e al contesto emotivo (Zovato et al., 2004). La nostra sensibilità a riguardo sta però cambiando e una generale accettazione per alcuni fenomeni «innaturali» che sono particolarmente ricorrenti in queste voci si sta affermando progressivamente.[17]

Un'analisi in termini di costi-benefici

Come anticipavo nell'introduzione, le occasioni che abbiamo oggi di fruire di una voce sintetica sono innumerevoli. Ma ancor di più, s'è detto, beneficiano di queste voci i non vedenti e gli ipovedenti, come accade, ad esempio, grazie ai lettori di schermo. Molti servizi pubblici e molta editoria se ne servono, ricorrendo a «voci», cioè timbri vocali che sono facilmente riconoscibili in diverse piattaforme.[18]

Molto lavoro un tempo richiesto ai dicitori di professione per leggere un comunicato o il testo di un racconto è ora spesso affidato a una di queste voci (o a un software che mette a disposizione un portfolio che può includere decine di voci in decine di lingue diverse).[19]

In molti casi, come ad esempio nelle applicazioni per navigatori satellitari, sarebbe impensabile affidare a un professionista della voce la generazione di tutti gli enunciati possibili (a causa dell'elevato numero di toponimi contenuti nel «vocabolario», che si potrebbero combinare in milioni di sequenze diverse). In quel caso, l'avvento delle voci sintetiche è stato addirittura condizione determinante per uno sviluppo a tutto campo della stessa tecnologia. I risultati possono non essere sempre ottimali, ma tant'è.

Invece, ancora oggi si discute sull'opportunità di ricorrere a una voce sintetica per audiodescrivere i contenuti di un prodotto multimediale o audio-visivo e renderlo accessibile: la vocalizzazione automatica consente tempi di lavorazione che risultano nettamente minori e che permettono di audiodescrivere una quantità decisamente

di professione, in grado di assicurare, più che una dizione standard, un buon controllo degli aspetti ritmico-intonativi del parlato, condizione per una sintesi più convincente. Oggi le possibilità sono diverse e si offrono a chiunque desideri sviluppare una voce *custom*. Un'azienda padovana sta sviluppando un sistema di sintesi vocale personalizzato che offre il mantenimento del timbro caratteristico del donatore. Si tratta di una tecnologia utile ad esempio ai pazienti di SLA che stanno perdendo la capacità di parlare e che grazie a questo sistema «potranno conservare per sempre il loro peculiare timbro e intonazione grazie alla registrazione di poche decine di frasi» (vedi in sitografia MIVOQ). Creato il modello vocale sulla base di un numero sorprendentemente contenuto di frasi, l'utente che ad esempio perda l'uso della parola potrà continuare a usufruire della propria voce grazie a un dispositivo elettronico portatile dotato di *TTS*.

[17] L'esperienza e l'abitudine dell'ascoltatore con le voci sintetiche e con gli ambienti di valutazione aveva inizialmente un profondo effetto sull'esito dei test, ma — con la progressiva diffusione di queste voci — la loro graduale accettazione ha migliorato la performance dei sistemi e ha persino generato un «gusto» per le loro caratteristiche e una sensibilità per i nuovi «accenti» prodotti (in Italia si pensi al successo delle imitazioni delle segreterie vocali da parte di artisti televisivi come Virginia Raffaele).

[18] Di alcune di queste si può fruire gratuitamente su diversi siti web (vedi sitografia). Tra le voci commerciali italiane ricordiamo, tuttavia, almeno quelle di Apple, Google o Amazon, che hanno accolto e integrato nei loro prodotti i moduli di sintesi sviluppati da ricercatori di diversi enti e centri di ricerca. Pensiamo ad esempio a Festival e Festvox *(Carnegie Mellon University's speech group)* o ai prodotti dei nostri Loquendo (a Torino, ora Nuance) e ITC-IRST (a Trento, ora FBK – Fondazione Bruno Kessler). Per indicazioni più precise a questo riguardo si veda Cosi (2003) e si faccia riferimento ai materiali pubblicati dall'AISV-ISCA (Associazione Italiana di Scienze della Voce – International Speech and Communication Association).

[19] Alcuni editori di testi scolastici mettono ormai a disposizione dei loro utenti (che, ad esempio, presentino Bisogni Educativi Speciali, BES) i file con la sintesi vocale del loro contenuto.

maggiore di prodotti. La voce sintetica è, inoltre, versatile nell'audiodescrizione di collezioni di testi variabili, che richiedono un costante aggiornamento (come ad esempio quelli presenti in un sito web).[20]

Rispetto a una voce naturale, i detrattori di questi servizi obiettano che la voce possa risultare talvolta innaturale, noiosa, senza emozioni. Inoltre, alcune di queste voci possono «incepparsi» nella pronuncia di sigle, acronimi e nomi propri.[21]

Tuttavia, considerevoli progressi sono attualmente attesi sul piano tecnologico; è, inoltre, auspicabile che cresca la sensibilità per la formazione di professionisti in grado di scrivere testi ottimizzati per queste applicazioni.[22]

Nel caso di uso di voci sintetiche per le audiodescrizioni (a condizione che la sintesi adottata sia di alta qualità), oltre ai vantaggi presentati sopra, ci sarebbe anche la maggior facilità, per il fruitore, di discernere la voce sintetica dalle voci originali del prodotto (attori, doppiatori e dell'eventuale voce narrante) e dalle voci fuori campo che possono costellare il sonoro dell'audiovisivo.

Per renderle efficaci occorre, però, sviluppare competenze «dedicate», di specialisti che sappiano redigere testi in grado d'inframmezzarsi in modo non invasivo ai dialoghi e agli eventi salienti dell'opera (Arma, 2014) e aggirare l'ostacolo di nomi insoliti e parole straniere la cui grafia potrebbe trarre in inganno il dispositivo *TTS*.

Conclusioni

Con questo contributo ho passato rapidamente in rassegna alcuni risvolti del diffondersi nella società di sistemi automatici in grado di scambiare informazioni linguistiche con gli esseri umani.

[20] Anche l'Università degli Studi di Torino (UniTO) si è recentemente dotata di un servizio simile. Da mercoledì 25 maggio 2016, «[I]l portale di Ateneo offre un nuovo servizio di vocalizzazione on line. Unito.it si trasforma così in un sito parlante, utile per chi ha problemi di vista o soffre di disturbi della lettura (come la dislessia), non è ancora in grado di leggere bene la nostra lingua, preferisce ascoltare anziché leggere, studia e vuole avvantaggiarsi del rinforzo mnemonico dovuto alla lettura». Il testo delle pagine web viene letto in tempo reale da una voce *TTS* di alta qualità e il prodotto sonoro può essere esportato in formato mp3. Il servizio, basato su *ReadSpeaker* (vedi sitografia), funziona su tutti i dispositivi, con qualsiasi browser o sistema operativo. Per queste notizie, in parte disponibili nell'annuncio sul portale di Ateneo, ringrazio anche Elisa Bernardi e lo staff dell'Università di Torino, che ha vegliato sulla buona riuscita del servizio di vocalizzazione online e ha dedicato un certo tempo a rivelarmi alcuni risvolti operativi.

[21] Si pensi alla «pronuncia» da parte di alcuni navigatori GPS degli odonimi francesi e tedeschi di località della Valle d'Aosta e dell'Alto-Adige. Qualcuno ricorderà anche le difficoltà a rendere adeguatamente nomi propri, sigle e indirizzi web nelle letture radiofoniche di Golem (qualche anno fa). Il ricorso a *Readspeaker* consente invece al portale di UniTO (vedi numero precedente) di superare brillantemente alcuni potenziali impedimenti dello scritto: è sorprendente, ad esempio, la resa corretta di «A tale scopo, si rammenta che l'art. 54 del D.Lgs. 30 marzo 2001, n. 165, introdotto dall'art. 1 comma 51 della Legge 190/2012», in cui «art.» è riconosciuto e generato come «articolo», «n.» come «numero» e «D.Lgs.» come «decreto legislativo». Similmente, le voci di IVONA del servizio web di audiodescrizione *TellMeWhat*, così come quelle di molti servizi complementari, risultano di solito abbastanza robuste a questo riguardo (a condizione di saperle «usare» adeguatamente).

[22] Questi servizi risultano quindi efficaci e convincenti, a condizione di suggerire una grafia che renda meno (o, possibilmente, non) ambigua la pronuncia di parole inconsuete. Al momento, il servizio di vocalizzazione di UniTO incontra difficoltà nella resa di «Ludovico di Savoia-Acaia» (http://www.unito.it/ateneo/chi-siamo/storia) per via della generazione di un «Acaia» con accento su una /i/ che non è realmente presente nella pronuncia italiana di questo nome. Non si sarebbe incorsi in questo problema se si fosse adottata una grafia meno ambigua (come quella usata dappertutto nel sito per indicare, ad esempio, l'«Aula Principe d'Acaja»).

Sorvolando i numerosi campi in cui queste applicazioni si stanno affermando in modo considerevole, mi sono soffermato su alcuni aspetti che riguardano il loro uso nelle tecnologie assistive e nel campo dell'accessibilità (con rimandi ad altri interessanti contributi di questo volume). Sebbene ancora oggi molti fruitori non siano soddisfatti della qualità raggiunta, è indubbio il servizio che i sistemi *TTS* hanno reso e stanno rendendo in moltissimi casi alle persone che ne beneficiano.

In conclusione, con un occhio attento ai progressi degli agenti assistivi che ci affiancheranno sempre di più in un futuro non più lontano, propongo la lettura di un passaggio emblematico delle conclusioni di Sandri (1985), scritto nel linguaggio di quei decenni: «Di notevole interesse sociale sono applicazioni di sistemi di uscita vocale nel settore biomedico per aiutare a risolvere problemi di handicap verbale e visivo: per esempio, in collegamento con un lettore ottico in grado di decodificare i caratteri scritti sulla pagina, un sistema di sintesi potrebbe "leggere" effettivamente brani di libri e giornali per i non vedenti» (Sandri, 1985, p. 80).

Quella che nelle previsioni di Sandri poteva sembrare una condizione difficile da conquistare è oggi stata raggiunta e oltrepassata, con ausili che hanno permesso di superare la complessità tecnica che allora si poteva riconoscere in alcuni passaggi. È il caso di dire, anche in questo caso, che la realtà ha superato le previsioni.

Strengths and weaknesses of synthetic voices in the field of assistive technology

Abstract

Synthetic voices are nowadays widespread in much-diversified uses and applications. In fact, the quality of speech synthesis has reached an acceptability that makes artificial voices sometimes indistinguishable from natural ones.

After a brief examination of the main advances in speech synthesis and human-machine interaction, this contribution offers an analysis in terms of costs and benefits of its use in some application environments, such as GPS navigation systems, home automation, text-to-speech systems, audio-books, audio-guides and audio-descriptions.

Keywords

Voice in applications, Speech synthesis, Man-machine interaction, Conversational agents, Audio-description.

Autore per corrispondenza

Antonio Romano
Università degli Studi di Torino
Dipartimento di Lingue e Letterature Straniere e Culture Moderne
Via Verdi, 10
10124 Torino
E-mail: antonio.romano@unito.it

Bibliografia

Arma V. (2014), *L'audiodescrizione: Stato dell'arte e prospettive di mercato in Italia,* in E. Perego (a cura di), *L'audiodescrizione filmica per i ciechi e gli ipovedenti,* Trieste, EUT, pp. 59-71.

Bailly G., Benoît C. e Sawallis T. (1992), *Talking machines: Theories, models, and design,* Amsterdam, Elsevier.

Bessière P., Boë L.-J. e Franceschini N. (2002), *250 ans après Vaucanson, les robots de l'an 2000,* Grenoble, Institut de la communication parlée.

Biondi T. (2012), *La narrazione al cinema: dal pensiero narrativo alla rappresentazione filmica,* Roma, Meti.

Cheyer A., Park J. e Giuli R. (2005), *IRIS: Integrate. Relate. Infer. Share. Proceedings of the ISWC 2005* (*Workshop on The Semantic Desktop - Next Generation Information Management & Collaboration Infrastructure, Galway, Irlanda, 6 novembre 2005),* http://ceur-ws.org/Vol-175/17_park_iris_final.pdf

Corbella M. (2015). *Narrazione e rappresentazione nella popular music e nel cinema all'incrocio tra voce e suono sintetizzato,* in A. Romano, M. Rivoira e I. Meandri (a cura di) (2015), *Aspetti prosodici e testuali del raccontare: Dalla letteratura orale al parlato dei media. Atti del X Convegno Nazionale dell'AISV,* Alessandria, Dell'Orso.

Cosi P. (2003), *Sintesi della voce e agenti parlanti,* «I quaderni di Telema», n. 21, pp. 74-78, http://www2.pd.istc.cnr.it/Papers/PieroCosi/cp-TELEMA2003.pdf

Fissore L., Laface P., Micca G. e Pieraccini R. (1992), *Performance of a Speaker-Independent Continuous Speech Recognizer,* in P. Laface e R. De Mori (a cura di), *Speech recognition and understanding: Recent Advances, Trends, and Applications,* Berlin, Springer, pp. 171-179.

Giannini A. e Pettorino M. (1999), *Le teste parlanti,* Palermo, Sellerio.

Kaszczuk M. e Osowski L. (2006), *Evaluating Ivona Speech Synthesis System for Blizzard Challenge 2006. Proceedings of Blizzard Challenge 2006 Workshop (Interspeech 2006 – ICSLP, Pittsburgh, PA, 16 settembre 2006,* http://www.festvox.org/blizzard/blizzard2006.html

Kaszczuk M. e Osowski L. (2007), *The IVO Software Blizzard Challenge 2007 Entry: Improving IVONA Text-To-Speech. Proceedings of BLZ3 (Blizzard 2007 – Sixth ISCA Workshop on Speech Synthesis, Bonn, 25 agosto 2007),* http://www.festvox.org/blizzard/bc2007/blizzard_2007/blz3_010.html

Klatt D. (1987), *Review of text-to-speech conversion for English,* «Journal of the Acoustical Society of America», vol. 82, pp. 737-793, http://www.cs.indiana.edu/rhythmsp/ASA/Contents.html

Liu Y., Chawla N., Harper M., Shriberg E. e Stolcke A. (2006), *A study in machine learning from imbalanced data for sentence boundary detection in speech,* «Computer Speech and Language», vol. 20, n. 4, pp. 468-494.

Liu Y., Shriberg E., Stolcke A., Hillard D., Ostendorf M. e Harper M. (2006), *Enriching speech recognition with automatic detection of sentence boundaries and disfluencies,* «IEEE Trans. Audio, Speech and Language Processing», vol. 14, n. 5, pp. 1526-1540, https://pal.sri.com/publications/pal-2006/

Paoloni A. (2002), *La voce come elemento di identificazione della persona,* in A. De Dominicis (a cura di), *La voce come bene culturale,* Roma, Carocci, pp. 125-139.

Pieraccini R. (2012). *The voice in the machine: Building computers that understand speech,* Cambridge (MA), MIT Press.

Pieraccini R. et al. (1994), *A speech understanding system based on statistical representation of semantics: Proceeding of ICASSP'92 – IEEE International Conference on Acoustics, Speech and Signal Processing,* vol. 1, pp. 193-196.

Rabiner L.R. e Schafer R.W. (1978), *Digital Processing of Speech Signals,* Englewood Cliffs (NJ), Prentice-Hall.

Roe D.B. e Wilpon J.G. (1994), *Voice communication between humans and machines,* Washington, National Academy Press.

Romano A., Cesari U., Mignano M., Schindler O. e Vernero I. (2012), *Voice quality / La qualità della voce,* in A. Paoloni e M. Falcone (a cura di), *La voce nelle applicazioni: Atti dell'VIII Convegno dell'AISV,* Roma, Bulzoni.

Romano A., Rivoira M. e Meandri I. (a cura di) (2015), *Aspetti prosodici e testuali del raccontare: Dalla letteratura orale al parlato dei media. Atti del X Convegno Nazionale dell'AISV,* Alessandria, Dell'Orso.

Romero-Fresco P. (2011), *Subtitling through speech recognition: Respeaking,* Manchester, St Jerome.

Romito L. (2000), *Manuale di fonetica articolatoria, acustica e forense,* Cosenza, Centro Editoriale e Librario - Università degli Studi della Calabria.

Sandri S. (1985), *La sintesi automatica della lingua italiana,* «Le Scienze», vol. 199, pp. 68-80, http://download.kataweb.it/mediaweb/pdf/espresso/scienze/1985_199_6.pdf

van Santen J.P.H., Sproat R., Olive J. e Hirschberg J. (1997), *Progress in speech synthesis,* Berlin, Springer.

Saracco R. e Penza M. (1999), *Una macchina che parla,* «Le Scienze», vol. 375, pp. 96-101, http://download.kataweb.it/mediaweb/pdf/espresso/scienze/1999_375_6.pdf

Vair C., Colibro D., Castaldo F., Dalmasso E. e Laface P. (2007), *Loquendo – Politecnico di Torino's 2006 NIST Speaker Recognition Evaluation System: Proceedings of Interspeech 2007 (Anversa, Belgio, 27-31 agosto 2007),* pp. 1238-1241.

Viswanathan M. e Viswanathan M. (2005), *Measuring speech quality for text-to-speech systems: development and assessment of a modified mean opinion score (MOS) scale,* «Computer Speech and Language», vol. 19, n. 1, pp. 55-83.

Zavagna P. (2013), *La voce senz'anima: origine e storia del Vocoder,* «Musica/Tecnologia», vol. 7, pp. 27-63, http://www.fupress.net/index.php/mt/article/view/13206

Zovato E., Pacchiotti A., Quazza S. e Sandri S. (2004), *Towards emotional speech synthesis: a rule based approach: Proceedings of SSW5-2004 (Fifth ISCA ITRW on Speech Synthesis, Pittsburgh, USA, 14-16 giugno 2004),* pp. 219-220.

Sitografia

Associazione Italiana di Scienze della Voce: http://www.aisv.it/

DARPA (Defense Advanced Research Projects Agency of the USA), «Voices from DARPA»: http://www.darpa.mil/news-events/2016-09-26

«Free online TTS service with natural sounding voices»: http://www.fromtexttospeech.com/

ICBSA (Istituto per i Beni Sonori e Audiovisivi), «Le Macchine Parlanti»: http://www.icbsa.it/index.php?it/794/le-macchine-parlanti

IVONA Text to Speech «High quality, natural-sounding Text-to-Speech voices»: https://www.ivona.com/us/about-us/voice-portfolio/

JIBO – R. Pieraccini «JIBO's Conversational Technology»: https://www.youtube.com/watch?v=nSt5eZeWkGo

MIVOQ – TTS technology «Create your digital voice»: https://www.mivoq.it/

Nuance Communications «Dragon NaturallySpeaking»: http://www.nuance.com/for-individuals/by-product/dragon-for-pc/index.htm

Nuance Communications «Interactive Loquendo TTS Demo»: http://www.nuance.it/azienda/soluzione/soluzioni-assistenza-clienti/servizi-soluzioni/inbound/loquendo-small-business-bundle/interactive-tts-demo/index.htm

Oddcast Media technology «Text To Speech»: http://www.oddcast.com/home/demos/tts/tts_example.php?sitepal

Romano A. «Voce e artefatti» (Giornata mondiale della Voce – Università degli Studi di Torino, 16 aprile 2015): https://www.youtube.com/watch?v=5YLkdtvHl9Q

ReadSpeaker «The Power of Speech»: http://www.readspeaker.com/it/

«Tell me What» – Audio Description: https://tellmewhat.eu/