ETICHETTATURA FONETICA E TTS

Pier Luigi Salza *, Claudia Barolo*, Beata Dobrzynska^
* Loquendo S.p.A. – Torino; ^ CELI s.r.l. – Torino
{pierluigi.salza;claudia.barolo}@loquendo.com; beata@celi.it

1. SOMMARIO

Nell'articolo si illustra la gestione dell'etichettatura fonetica attuata all'interno del sistema di sintesi da testo LoquendoTTS® multilingue e multilocutore basato sulla tecnica di sintesi concatenativa per selezione di unità. L'etichettatura fonetica rappresenta uno strumento delicato e al tempo stesso essenziale per lo sviluppo del sistema. Il codice per la trascrizione, infatti, deve essere "universale", nel senso che deve permettere di rappresentare contemporaneamente, in maniera congruente e non ambigua, sia tutti i foni delle diverse lingue, quindi essere *language-independent*, sia tutte le varianti allofoniche dei singoli locutori, quindi essere *speaker-dependent*.

Un grande livello di dettaglio e di accuratezza fonetica è necessario per soddisfare la compatibilità con il segmentatore fonetico automatico dei *database* acustici registrati, mentre la non ambiguità nella rappresentazione dei foni tra lingue diverse è necessaria per attuare il *mapping* fonetico multilingue, ovvero la possibilità che qualsiasi "voce" del TTS, utilizzando il proprio sistema fonologico, pronunci parole o frasi "in prestito" da una qualsiasi delle altre lingue supportate dal sistema, e anche per consentire, infine, l'accesso a lessici fonetici esterni.

Ciò è possibile solo garantendo l'univocità del codice di trascrizione, attuata attraverso la corretta e rigorosa associazione di ciascun simbolo fonetico a tratti fonetico-articolatori (*features*) pertinenti, assimilabili a quelli dell'IPA. L'etichettatura fonetica nel sistema LoquendoTTS® si esplica a tre diversi livelli di granularità: livello di lingua o di parola: trascrizione larga o fonemica; livello di frase: trascrizione stretta; livello di voce: trascrizione allofonica. Il sistema LoquendoTTS®, attraverso questa accurata ed univoca etichettatura fonetica, è in grado di gestire una pluralità di codici: un codice interno, assimilabile a quello dell'IPA, il codice SAMPA e la sua estensione XSAMPA.

2. INTRODUZIONE

Il sistema LoquendoTTS® (Text-to-Speech) multilingue e multilocutore si basa sulla tecnica di sintesi concatenativa per selezione di unità non uniformi e contestuali, anche detta tecnica *corpus based*. Tale tecnica consiste nella possibilità di generare il segnale vocale a partire da un *database* di segmenti acustici registrati direttamente dalla voce naturale di uno *speaker* professionista nella sua lingua nativa. Il *database* contiene quindi i foni di una singola lingua e di un singolo parlatore. Le unità acustiche considerate, al contrario di quelle tradizionalmente utilizzate (singoli foni, coppie di foni o difoni, trifoni, demisillabe), sono di lunghezza non definita a priori (perciò non uniformi) e disponibili in differenti varianti prosodiche (dunque contestuali), distinte mediante informazioni aggiuntive di tipo sintattico-prosodico.

I sostanziali vantaggi dell'uso di tali unità sono: 1) l'**ampliamento della estensione** delle unità, che, includendo eventualmente anche parole intere o perfino parti di messaggi, riduce il numero delle giunzioni e quindi delle relative discontinuità; 2) la **contestualità**,

che diminuisce la necessità di modificare drasticamente la prosodia in fase di sintesi, evitando le conseguenti distorsioni del segnale vocale. Il risultato complessivo è che la conoscenza acustico-fonetica di una certa lingua rimane precostituita nelle unità acustiche stesse e il timbro della voce del donatore viene preservato praticamente inalterato, conservando gran parte delle sue caratteristiche naturali. La sintesi mediante questa tecnica ha una qualità acustica paragonabile a quella del parlato preregistrato.

In un sintetizzatore vocale per testi qualsiasi, la soluzione delineata comporta l'attuazione di un difficile compromesso tra estensione e contestualità delle unità acustiche (quindi il loro numero e la pluralità delle loro occorrenze) da un lato e le dimensioni del *database* acustico da registrare dall'altro, per ottenere una consistente copertura fonetica e prosodica della lingua.

Occorre definire, innanzi tutto, una fotografia statistica della lingua. Essa è realizzata mediante algoritmi in grado di analizzare una grande mole di testi trascritti foneticamente, estrarre tutte le sequenze fonetiche da 2 a "n" foni, differenziate in base alla posizione occupata nel sintagma, che è l'unità linguistica prescelta come riferimento per l'assegnazione dell'etichettatura sintattico-prosodica, ordinare infine le sequenze in base alla frequenza di occorrenza. La fotografia statistica del dominio è assunta come *target* di copertura. Si estrae quindi in modo automatico un sottoinsieme minimo di frasi da registrare, foneticamente dense e adatte a coprire il *target*. Le frasi, una volta registrate, vengono poi segmentate ed etichettate automaticamente da un segmentatore fonetico che allinea ciascuna forma d'onda con la corrispondente trascrizione fonetica. Il segmentatore necessita di un addestramento per ciascuno *speaker*, a partire da un piccolo insieme di frasi registrate e dalla loro corrispondente segmentazione manuale.

La sequenza di foni viene quindi automaticamente etichettata con i valori dei parametri acustico-prosodici intrinseci (F0, durata e intensità), estratti mediante algoritmi di analisi del segnale. Un algoritmo specifico marca ogni periodo di F0 (*pitch marker*). A ciascun fono viene infine associata l'etichetta con la categoria prosodica, che identifica il tipo di sintagma (conclusivo, sospensivo, interrogativo) e la posizione del fono nel sintagma cui appartiene. Questa classificazione ha lo scopo di permettere la selezione di varianti significativamente diverse di ciascuna possibile unità.

Il dizionario di unità acustiche contestuali non-uniformi è quindi costituito da un insieme di frasi registrate, spezzate in sintagmi, segmentate e annotate, da utilizzare in linea come serbatoio di elementi potenziali, la cui natura e dimensione non sono definite a priori, ma sono da definirsi *run-time* ad ogni selezione necessaria per generare, in fase di sintesi, un qualsiasi messaggio richiesto.

La selezione delle unità in fase di sintesi è attuata da un algoritmo che sfrutta in modo ottimale il materiale vocale immagazzinato, mirando all'individuazione delle unità migliori, cioè le sequenze più lunghe che meglio si combinano, avendo come obiettivo quello di minimizzare sia le differenze spettrali nei punti di giunzione (contestualità fonetica), sia le differenze tra i parametri acustico-prosodici intrinseci (contestualità prosodica). E' quest'ultima condizione a richiedere che almeno le unità più frequenti vengano coperte da più versioni, tra le quali poter scegliere, al momento della concatenazione, quella che, in quel determinato contesto, minimizza la necessità di modifiche dei parametri prosodici e quindi le conseguenti distorsioni.

Conoscenze linguistiche generali e specifiche di lingua sono perciò fondamentali per lo sviluppo di ogni nuova lingua che si chiede al TTS di sintetizzare. Il punto di partenza per lo sviluppo di tali conoscenze consiste nella definizione del *set* di foni occorrenti nella lingua in esame e delle regole di trascrizione grafema-fono. Si comprende quindi come la

gestione dell'etichettatura fonetica rappresenti uno strumento delicato e al tempo stesso essenziale per tutto il sistema.

Fondandosi sull'utilizzo della tecnica descritta e sul supporto di esperti consulenti madrelingua, sono state sinora sviluppate per il sistema LoquendoTTS® le lingue elencate nell'APPENDICE A. Per una descrizione maggiormente dettagliata si rimanda a (Quazza *et alii*, 2001)

In Figura 1 è riportato lo schema architetturale del sistema LoquendoTTS®, nel quale sono evidenziate alcune funzionalità del sistema in cui l'etichettatura fonetica riveste un ruolo particolarmente importante. Si tratta di: **Segmentatore automatico**, di cui si è parlato sopra; ***Mapping* fonetico** multilingue, ovvero la possibilità che qualsiasi "voce" del TTS possa pronunciare, utilizzando il proprio sistema fonologico, parole o frasi "in prestito" da altre lingue; **Ingresso fonetico**, ovvero la possibilità che il sistema legga una stringa scritta foneticamente (come previsto, ad es. dalla necessità di accedere ad un lessico esterno).

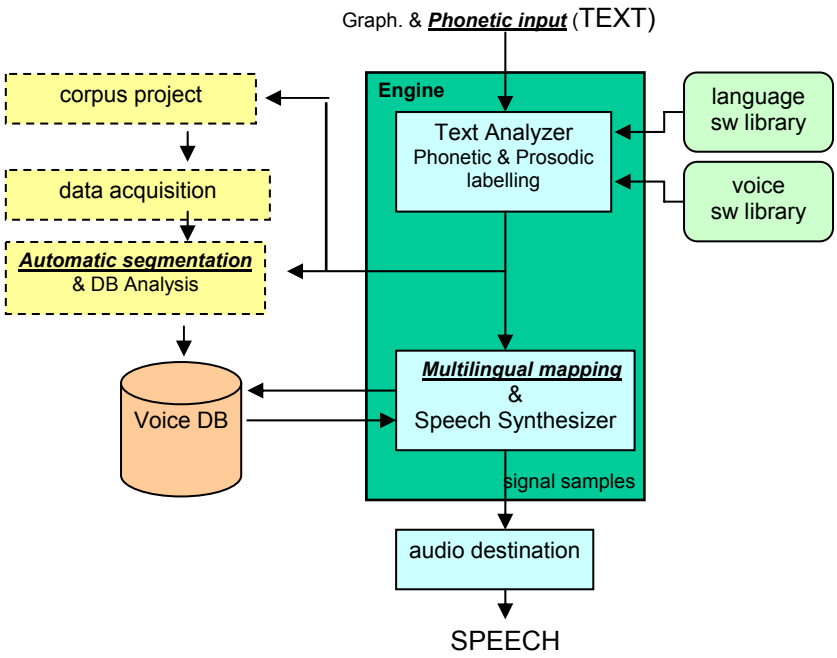

Figura 1. Schema architetturale del sistema LoquendoTTS®.

3. ETICHETTATURA FONETICA E TTS

Da quanto esposto risulta evidente la necessità di attuare, all'interno del sistema LoquendoTTS® multilingue e multilocutore, la gestione di un *set* di foni "potenzialmente

infinito". Il codice per la trascrizione, infatti, deve essere "universale", nel senso che deve permettere di rappresentare contemporaneamente, in maniera congruente e non ambigua, sia tutti i foni delle diverse lingue, quindi essere *language-independent*, sia tutte le varianti allofoniche dei singoli locutori, quindi essere *speaker-dependent*, consentendo il livello di dettaglio e di accuratezza fonetica atto a soddisfare la compatibilità con il segmentatore fonetico. L'ambiguità, nella rappresentazione dei foni, tra lingue diverse non è accettabile se si deve poter passare da una lingua all'altra (*mapping* fonetico). Ciò significa che la precisione del codice di trascrizione deve essere estremamente accurata, fondata su scelte giustificate foneticamente e pragmaticamente adattate all'applicazione.

La gestione dell'etichettatura fonetica nel sistema LoquendoTTS® si fonda sull'uso rigoroso di tratti fonetico-articolatori assimilabili a quelli dell'IPA (International Phonetic Association). Con "codice", infatti, si intende l'insieme del simbolo e dei tratti fonetico-articolatori (*features*) associati. L'elenco dei tratti utilizzati da LoquendoTTS® per la codifica fonetica è riportato nell'APPENDICE B. Attraverso un'etichettatura attuata mediante queste *features*, LoquendoTTS® gestisce, oltre ad un codice interno assimilabile al codice dell'IPA, il codice SAMPA (Speech Assessment Methods Phonetic Alphabet) (Wells, 1997), divenuto una sorta di "standard" a livello europeo (van den Heuvel, 2003), e il codice XSAMPA (eXtended SAMPA), versione estesa del SAMPA, che possiede la stessa struttura e quindi la stessa potenzialità di specificazione dell'IPA, proposto dallo stesso autore (si veda il sito www.phon.ucl.ac.uk/home/sampa). In Figura 2 sono riportati alcuni esempi di codifica di foni con il codice IPA, con il codice SAMPA e con quello XSAMPA.

IPA	SAMPA	XSAMPA	TRATTI ARTICOLATORI TTS
[ɐ̃]	[6~]	[6~]	VOWEL | RSDOPEN | CENTRAL | NASALIZED
[a͡ʊ]	[aU]	[aU]	VOWEL | OPEN | CENTRAL | LWRDCLOSE2 | BACK2 | ROUNDED2
[β̪]	[B]	[B\]	CONSONANT | APPROXIMANT | BILABIAL | VOICED

Figura 2. Alcuni esempi di codifica di foni con il codice IPA, con il codice SAMPA e con il codice XSAMPA. RSDOPEN: Rised Open; LWRDCLOSE: Lowered Close; 2: secondo segmento vocalico del dittongo.

L'etichettatura fonetica nel sistema LoquendoTTS® si esplica a tre diversi livelli di granularità, che vanno da quello più largo, fonemico, a quello più stretto, allofonico. A ciascun livello di trascrizione corrisponde un *set* di simboli più o meno ampio, con associati tratti fonetico-articolatori, espresso in forma tabellare, come in Figura 2.

Il primo livello è quello della **trascrizione di lingua**. Si tratta di una trascrizione grafema-fonema larga, con cui il testo in ingresso viene trascritto parola per parola. Tale livello di trascrizione fa riferimento ad un *set* di foni contenuto in una Tabella di Lingua. Questo livello di trascrizione deve assicurare la compatibilità con ingressi fonetici esterni (accesso a lessici trascritti foneticamente) e consentire la costruzione di lessici fonetici da parte dell'utente.

Il secondo livello di trascrizione è quello delle regole di riscrittura fonema-fonema. Tale livello di trascrizione è detto anche livello di **trascrizione di frase**. A questo livello vengono trattate le realizzazioni fonetico-articolatorie occorrenti nel contesto di frase

mediante regole di riscrittura che agiscono sia all'interno (occorrenze allofoniche di lingua) sia ai confini tra le parole (effetto dei fenomeni di coarticolazione), incluse le deaccentazioni. Si tratta di una trascrizione stretta, o fonetica (allofonica), che fa ancora riferimento ai foni contenuti nella Tabella di Lingua. Tale trascrizione deve garantire la coerenza con le effettive realizzazioni fonetico-articolatorie dipendenti dalle caratteristiche della lingua in esame a livello di frase e deve allo stesso tempo garantire la non–ambiguità tra le diverse lingue.

Infine, vi è un terzo livello di trascrizione, detto livello di **trascrizione di voce**, che consente l'inserzione di allofoni speaker-dependent, anche diversi per le differenti voci di una certa lingua. Si tratta di una trascrizione stretta, o fonetica (allofonica) idiosincratica, che agisce ancora in base a regole di riscrittura fonema-fonema, cui corrisponde un *set* di foni più ampio contenuto in una Tabella di Voce. Tale trascrizione deve garantire l'operatività del segmenatore fonetico automatico *voice dependent* e dell'algoritmo specifico che marca ogni periodo di F0 (*pitch marker*). A tale livello di trascrizione, l'esatta attribuzione del tratto fonetico-articolatorio *voiced/voiceless* (e *devoiced*) è estremamente importante, sia per l'addestramento del segmentatore sia per il funzionamento del *pitch marker*.

Nella Figura 3 è illustrato un esempio di trascrizione su più livelli dell'enunciato in lingua Inglese britannico "*pests became*".

Trascrizione di lingua (o di parola): Tabella di Lingua

- IPA: [ˈp ɛ s t͡s b ɪ ˌk e͡ɪ m]
- SAMPA: ["p E s ts b I 'k eI m]
- XSAMPA: ["p E s ts b I 'k eI m]

Trascrizione di frase: Tabella di Lingua

- IPA: [ˈpʰ ɛ s t͡s b ɪ ˌkʰ e͡ɪ m]
- SAMPA: ["p E s ts b I 'k eI m]
- XSAMPA: ["p_h E s ts b I 'k_h eI m]

Trascrizione di voce: Tabella di Voce (voce Simon)

- IPA: [ˈpʰ ɛ s t͡s b̥ ɪ ˌkʰ e͡ɪ m]
- SAMPA: ["p E s ts b I 'k eI m]
- XSAMPA: ["p_h E s ts b_0 I 'k_h eI m]

Figura 3. Esempio di trascrizione multilivello dell'enunciato: "*pests became*", in Inglese britannico. Nota: in Inglese britannico la plosiva *voiceless*, in determinati contesti, può essere aspirata oppure *unreleased*; inoltre, sono frequenti le realizzazioni *speaker dependent devoiced* delle plosive *voiced*, come [b], in contesto *voiceless* (ad es. con la voce Simon).

Dall'esempio mostrato si può osservare come il codice SAMPA non consenta il grado di accuratezza necessario per la rappresentazione degli allofoni, tanto a livello di frase quanto a livello di voce.

4. IMPORTANZA DELL'ACCURATEZZA FONETICA

Come già evidenziato nello schema della Figura 1, per alcune funzionalità del sistema LoquendoTTS® l'etichettatura fonetica è cruciale. Ora mostreremo qualche maggiore dettaglio relativamente a: **Segmentatore fonetico** automatico, ***Mapping* fonetico** multilingue su testi "mixed language" e *mapping* fonetico con altro codice (SAMPA) per **Ingresso fonetico** al TTS.

4.1 Segmentatore fonetico automatico

Il Segmentatore fonetico automatico è un software che, utilizzando il segnale del parlato registrato e la sua trascrizione fonetica, realizza automaticamente la segmentazione del *database* acustico in foni, difoni e demifoni (mezzi foni), i "mattoncini" più piccoli eventualmente utilizzabili durante la sintesi per la generazione del messaggio, quando non sono reperibili sequenze più lunghe. Il livello di trascrizione necessario è quello di voce. Il software si basa su modelli markoviani (HMM, Hidden Markov Model) e regole (Brugnara *et alii*, 1993). La trascrizione fornisce non solo l'etichettatura fonetica, ma anche l'etichettatura prosodico-posizionale della sillaba cui il fono appartiene, ad es.: prima tonica, tonica interna, atona, ultima tonica, sillaba finale di sintagma/enunciato. Il Segmentatore fonetico automatico opera sulla singola voce mediante previo addestramento. Tutte le peculiarità *speaker-dependent* devono essere rilevate dalla trascrizione e fornite al al segmentatore. I foni devono essere etichettati sul *database* acustico così come effettivamente pronunciati dal parlatore perciò ogni discrepanza tra la trascrizione e la pronuncia effettiva va individuata e corretta. Un'analisi acustico fonetica manuale viene quindi effettuata mediante ascolto e visualizzazione del segnale. Si tratta, per la gran parte, di fenomeni di sonorizzazione e desonorizzazione delle consonanti o riduzione delle vocali. Un esempio è dato dalla desonorizzazione delle Occlusive e Approssimanti in Castigliano, da parte della voce Jorge: [b d̪ g β̞ ð̞ ɰ] -> [b̥ d̪̊ g̊ β̊ ð̊ ɰ̊]. In Figura 4 è riportato un esempio di trascrizione a livello di voce (voce Jorge) della parola "*amaba*" in Castigliano.

- IPA: [a ˈm a β̞ a]
- SAMPA: [a "m a B a]
- XSAMPA: [a "m a B_0 a]

Figura 4. Esempio di trascrizione a livello di voce, con i codici IPA, SAMPA e XSAMPA, della parola "*amaba*" in Castigliano (voce Jorge).

4.2 Mapping fonetico su testi " mixed-language"

La conversione testo-voce è intrinsecamente language-dependent, tuttavia le applicazioni del TTS, a partire dai testi reali di una molteplicità di applicazioni, includono sempre più spesso parole o intere frasi che sono prese in prestito da altre lingue. In questi casi i testi sono detti "*mixed language*". L'obiettivo è ottenere una pronuncia dei termini stranieri mantenendo la stessa voce. Trattandosi di lettura di parole o frasi in L2 presenti in un testo della lingua corrente L1, la pronuncia attesa non è esattamente quella prevista in L2, bensì una sua versione approssimata. Solitamente, infatti, il parlante di L1 tende a pronunciare questi "inserti" di L2 mantenendo il più possibile il proprio sistema fonologico.

Nel sistema LoquendoTTS® multilingue e multilocutore si è realizzato un algoritmo che abilita ciascuna voce a parlare in tutte le lingue supportate dal sistema, utilizzando il proprio sistema fonologico e basandosi su un principio detto *mapping* fonetico multilingue (Badino *et alii*, 2004). Tale principio sfrutta l'accurata e non ambigua etichettatura di foni e allofoni delle varie lingue/voci, più sopra descritta, effettuata mediante attribuzione di simboli associati a una stringa di tratti fonetico-articolatori. Il criterio fondamentale per la scelta del fono di L1 che più si avvicina a quello di L2 da pronunciare e mancante in L1 è che due foni possono essere giudicati simili quando essi sono caratterizzati da tratti fonetico-articolatori simili. In sostanza i foni di L2 mancanti in L1 sono realizzati, con una certa approssimazione, scegliendo i foni più simili nel *database* della lingua corrente L1.

La principale difficoltà del *mapping* è rappresentata dalla definizione del peso di ciascun tratto fonetico-articolatorio all'interno di una funzione di stima della similarità (o distanza) tra tratti. Particolarmente difficile è calcolare il grado di affinità tra i valori dei tratti "non binari", sia quelli che si riferiscono al modo e luogo di articolazione delle consonanti sia quelli che specificano la posizione delle vocali all'interno del quadrilatero vocalico. Un esempio molto semplice e comune di *mapping* fonetico è la pronuncia dei foni [p^h] e [$p^\neg$] da parte di una voce italiana che parla Inglese: in entrambi i casi il fono scelto sarà [p]. In Figura 5 e in Figura 6 si riportano due esempi di *mapping*, rispettivamente tra Italiano e Inglese britannico e tra Italiano e Inglese americano, riguardanti la pronuncia della stessa frase: "*Her party was a success*".

- IPA: [h ɜː ˈpʰ aː t̪ i w ə z ə s ʌ kˀ ˈs ɛ s]
- SAMPA: [h 3: "p A: t i w @ z @ s V k "s E s]
- XSAMPA: [h 3: "p_h A: 4_d i w @ z @ s V k_} "s E s]

- IPA: [e ˈp a t̪ i w e z e s e k ˈs ɛ s]
- SAMPA: [e "p a t i w e z e s e k "s E s]
- XSAMPA: [e "p a t i w e z e s e k "s E s]

Figura 5. Esempio di *mapping* tra Italiano e Inglese britannico per la frase "*Her party was a success*", trascritto in IPA, in SAMPA e in XSAMPA. Prima versione: trascrizione a livello di frase in Inglese britannico; seconda versione: trascrizione del *mapping* tra l'Italiano e l'Inglese britannico (voce Paola).

- IPA: [h ɜʷː ˈpʰ aː ɹ t̪ i w ə z ə s ʌ kˀ ˈs ɛ s]
- SAMPA: [h @ r: "p A: r t i w @ z @ s V k "s E s]
- XSAMPA: [h 3` : " p_h A: r\ 4_d i w @ z @ s V k_} "s E s]

- IPA: [e r ˈp a r t̪ i w e z e s e k ˈs ɛ s]
- SAMPA: [e r "p a r t i w e z e s e k "s E s]
- XSAMPA: [e r "p a r t i w e z e s e k "s E s]

Figura 6. Esempio di *mapping* tra Italiano e Inglese americano per la frase "*Her party was a success*", trascritto in IPA, in SAMPA e in XSAMPA. Prima versione: trascrizione a livello di frase in Inglese americano; seconda versione: trascrizione del *mapping* tra l'Italiano e l'Inglese americano (voce Paola).

4.3 Mapping fonetico con altro codice: SAMPA

Al fine di conseguire una corretta pronuncia è spesso necessario ricorrere a lessici fonetici esterni. E' questo il caso, in quasi tutte le lingue, di molti nomi propri, in particolare cognomi e toponimi, che fanno eccezione alle normali regole di pronuncia. I lessici fonetici utilizzano normalmente trascrizioni con il codice SAMPA, mentre la versione estesa XSAMPA non è praticamente ancora utilizzata.

D'altro canto, è necessario rendere disponibile, oltre all'uscita acustica del TTS, anche un'uscita scritta foneticamente in un codice facilmente leggibile e immediatamente utilizzabile in determinate applicazioni.

Queste due funzioni del sistema LoquendoTTS® sono indicate come *input* e *output* fonetico del TTS. Per ciascuna lingua/voce sono state perciò create due ulteriori tabelle, una in codice SAMPA e una in codice XSAMPA.

Come già accennato in precedenza, il codice SAMPA è un codice per trascrizione larga o fonemica, la cui "grana" non è sufficientemente accurata per una trascrizione fonetica allofonica. Spesso, le trascrizioni SAMPA dei lessici esterni risultano affette da diverse criticità, in particolare: l'incongruenza e la sottospecificazione. L'**incongruenza** ha a che fare con un uso non coerente dei simboli di alcuni foni e con i diacritici. Ad esempio, nei lessici, spesso viene rispettata la non ambiguità intralingua ma non quella interlingue. Ai fini dell'intelligibilità e della fruibilità del singolo lessico (che ovviamente è monolingua) ciò può risultare sufficiente. Per il TTS multilingue ciò rappresenta invece un problema. Non è raro trovare lessici dell'Inglese britannico trascritti in SAMPA in cui l'approssimante alveolare [ɹ] è trascritta semplicemente [r], come nell'esempio "*sorry"* illustrato nella Figura 7.

- IPA: [ˈs ɔːɹ i]
- SAMPA: ["s O: r i]
- XSAMPA: ["s O: r\ i]

Figura 7. Esempio di trascrizione a livello di voce, con i codici IPA, SAMPA e XSAMPA, della parola "*sorry*" in Inglese britannico (voce Simon)

Quanto alla **sottospecificazione** rispetto alle corrispondenti trascrizioni IPA, capita, ad esempio, che in un lessico Olandese si trovino utilizzati i simboli SAMPA [G], [@ n], [r], [l] in luogo dei simboli IPA [x], [ə̃], [ərˑ], [ɬ]. Nella Figura 8 sono illustrati i casi [@ n] e [r] nell'enunciato in lingua Olandese "*Baarzenstraat*".

- IPA: [ˈb aərˑ z ə̃ ˈs t ʁ a t]
- SAMPA: ["b a: r z @ n s t r a: t]
- XSAMPA: ["b a: @` z @~ n s t R a: t]

Figura 8. Esempio di trascrizione a livello di voce, con i codici IPA, SAMPA e XSAMPA, della parola "*Baarzenstraat*" in Olandese (voce Willem).

Si incontrano anche frequenti esempi di trascrizioni SAMPA in cui si evidenzia la neutralizzazione di alcune vocali con lo schwa, come mostrato nella Figura 9 per l'enunciato in lingua inglese americano "*east Los Angeles*", e nella Figura 10 per l'enunciato in lingua tedesca "*Erle Straße*".

- IPA: [iː s t l oʊ ˈs æ n d͡ʒ ə l ɪ s]
- SAMPA: [i s t l O " s { n dZ @ l @ s]
- XSAMPA: [i: s t l oU " s { n dZ @ l I s]

Figura 9. Esempio di trascrizione a livello di voce, con i codici IPA, SAMPA e XSAMPA, dell'enunciato "*east Los Angeles*" in lingua inglese americano (voce Dave).

- IPA: [ɛ ɐ l eː ˈʃ t ʁ aː s ə]
- SAMPA: [E 6 l @ "S t R a: s @]
- XSAMPA: [E 6 l e: "S t R a: s @]

Figura 10. Esempio di trascrizione a livello di voce, con i codici IPA, SAMPA e XSAMPA, dell'enunciato "*Erle Straße*" in lingua tedesca (voce Stefan).

In tutti questi casi, il *mapping* fonetico consente di selezionare nel *set* di foni di voce quello effettivamente realizzato dallo speaker.

5. CONCLUSIONI

L'accurata etichettatura fonetica adottata nel sistema di sintesi multilingue e multilocutore LoquendoTTS®, fondata sull'uso rigoroso di tratti fonetico-articolatori assimilabili a quelli dell'IPA, consente una gestione qualitativamente elevata della segmentazione fonetica automatica dei *database* acustici registrati e dell'accesso a lessici fonetici esterni. Inoltre essa fornisce allo stesso tempo una soddisfacente soluzione alla possibilità che qualsiasi voce, attraverso un algoritmo di *mapping* fonetico multilingue, possa leggere parole straniere in una qualsiasi delle lingue supportate dal sistema.

BIBLIOGRAFIA

Brugnara, F., Falavigna, D., & Omologo, M. (1993), Automatic Segmentation and Labelling of Speech based on Hidden Markov Models, *Speech Communication,*12, 357-370.

Badino, L., Barolo C., Quazza S. (2004), A general approach to TTS reading of mixed-language texts, in *Proceedings of ICSLP 2004*.

Quazza, S., Donetti, L., Moisa, L. & Salza, P. L. (2001), ACTOR®: a multilingual unit-selection speech synthesis system, in *Proceedings of 4th ISCA Tutorial and Research Workshop on Speech Synthesis 2001*, Perthshire, Scotland, August 29 - September 1.

Van den Heuvel, H., Choukri, K., Höge, H, Maegaard, B., Odijk, J. & Mapelli V. (2003), Quality control of language resources at ELRA, in *Proceedings of EUROSPEECH 2003*, Geneva, Switzerland.

Wells, J.C., (1997), SAMPA computer readable phonetic alphabet, in Handbook of Standards and Resources for Spoken Language Systems (D. Gibbon, R. Moore, & R.Winski, editors), 1997, Berlin and New York: Mouton de Gruyter, Part IV, section B.

APPENDICE A.

Lingue attualmente (03-08) sviluppate per LoquendoTTS®.

American	French	Portuguese
Am. Span. (Columbian)	French Canadian	Swedish
Argentine	Galician	Turkish
Brazilian	German	Valencian
British	Greek	
Castilian	Italian	
Catalan	Mandarin Chinese	
Chilean	Mexican	
Dutch	Polish	

Lingue in fase di sviluppo:

Russian, Danish, Finnish

APPENDICE B.

Tratti fonetico-articolatori (*features*) utilizzati da LoquendoTTS® per l'etichettatura fonetica, dove: LWRDCLOSE: Lowered Close; RSDOPEN: Raised Open; 2: secondo segmento vocalico del dittongo; LTRLTAPFLAP: Lateral TapFlap; LTRLFRICATIVE: Lateral Fricative; LTRLAPPROXIMANT: Lateral Approximant.

COMMON FEATURES:

Class: WOWEL, CONSONANT, PAUSE, NONPHONEME
Length: SHORT, LONG, EXTRASHORT, EXTRALONG
Voicing: VOICED, VOICELESS

COMMON DIACRITICS:

NASALIZED, DEVOICED, BREATHY, CREAKY

VOWEL FEATURES:

Rounding: NONROUNDED, ROUNDED
Heigth: CLOSE, LWRDCLOSE, CLOSEMID, MID, OPENMID, RSDOPEN, OPEN
Backness: FRONT, CENTRAL, BACK
Rounding2: NONROUNDED2, ROUNDED2
Heigth2: come sopra
Backness2: come sopra
Stress: UNSTRESSED, STRESSED
Tone: T1, T2, T3, T4, T5, T6, T7

VOWEL DIACRITICS:

NONSYLLABIC, RHOTICIZED, MOREROUNDED, LESSROUNDED, ADVANCED, RETRACTED

CONSONANT FEATURES:

Manner: PLOSIVE, NASAL, TRILL, TAPFLAP, LTRLTAPFLAP, FRICATIVE, AFFRICATE, LTRLFRICATIVE, APPROXIMANT, LTRLAPPROXIMANT, CLICK, IMPLOSIVE
Place: BILABIAL, LABIODENTAL, DENTAL, ALVEOLAR, POSTALVEOLAR, (RETROFLEX), PALATAL, VELAR, UVULAR, PHARINGEAL, GLOTTAL, EPIGLOTTAL

CONSONANT DIACRITICS:

ASPIRATED, UNRELEASED, NASALRELEASE, LATERALRELEASE, SYLLABIC, WITHVOICING, LINGUOLABIAL, LABIALIZED, PALATALIZED, VELARIZED, PHARINGEALIZED, LAMINAL, APICAL, RAISED, LOWERED