Sintesi vocale su dispositivi mobili

01 Ottobre 2015

http://www.mivoq.it/
info@mivoq.it MIVOQ S.R.L. +39 049 0998335

Indice

1 Obiettivo del progetto

L'obiettivo di questo progetto è di sperimentare e rendere disponibili su dispositivi mobili nuove funzionalità di sintesi vocale da testo (TTS: Text-To-Speech), normalmente non presenti su questi dispositivi, come la possibilità di applicare effetti alle voci o di poter utilizzare la voce degli utenti.

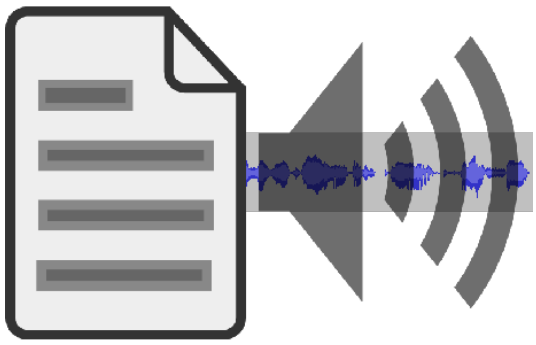

Figura 1: Flexible and Adaptive Text-To-Speech
`http://lab.mediafi.org/try-flexibleandadaptivetexttospeech.html`

La sintesi vocale è quella tecnologia che permette la conversione di un qualsiasi file di testo in un file sonoro. Negli ultimi anni si è assistito ad un rapido diffondersi di questo tipo di tecnologia in numerosi ambiti (per es.: voci guida dei navigatori satellitari, annunci dei mezzi di trasporto pubblico, centralini telefonici, lettori di messaggi) e al suo affermarsi come una delle interfacce di fruizione abituale per tutte quelle applicazioni in cui è impedito, o comunque limitato, l'utilizzo della vista (per es.: durante la guida).

Figura 2: Sygic GPS navigation and maps

La forte diffusione (funzionalità di sintesi vocale sono integrate in ogni smartphone, nella maggior parte dei navigatori GPS, in gran parte dei centralini telefonici, in gran parte dei lettori di ebook, in alcuni sistemi di annunci automatici) e l'affermarsi del successo di questa tecnologia in alcuni ambiti specifici (come per esempio nei navigatori GPS per auto o nella telefonia) hanno portato alla nascita di numerosi standard, che i vari motori (engine) di sintesi vocale possono implementare e al consolidamento delle funzionalità esposte. Le aziende del settore si sono quindi concentrate sul miglioramento della qualità timbrica e acustica delle voci generate, piuttosto che sulla ricerca di nuove funzionalità e potenzialità.

Grazie ai risultati recenti della ricerca, si sono rese disponibili nuove tecnologie in grado di rendere percorribili nuove strade (come la creazione di applicazioni di sintesi vocale per il canto, la possibilità per l'utente di creare la propria voce, o di alterarne lo stile per realizzare effetti particolari simulando emozioni o altre caratteristiche) e numerose aziende stanno investendo in questo campo. Contemporaneamente i centri di ricerca hanno cominciato a produrre e a fornire gratuitamente voci dalla qualità sempre migliore.

Lo scopo di questo progetto è di sfruttare alcune potenzialità aggiuntive fornite da questi sistemi, ed in particolare la possibilità di applicare effetti e di utilizzare la voce degli utenti stessi, per realizzare applicazioni innovative (per es.: lettura di SMS con la voce del mittente, personalizzazione della voce per il navigatore, ...) che vadano oltre quello che si può realizzare con sistemi standard, gettando le basi per nuove applicazioni e per l'introduzione di nuove funzionalità nella vita di tutti i giorni.

2 Software di sintesi vocale su dispositivi mobili

Il progetto prevede la realizzazione di un'applicazione per dispositivi mobili (smartphone e tablet), che sfrutti appieno le potenzialità offerte dal motore di sintesi opensource "Flexible and Adaptive Text To Speech" (FA-TTS) ed in particolare:

- possibilità di applicare effetti e stili alle voci (per es.: partendo da una voce tradizionale, realizzare una voce di bambino, una voce robotica o quella di un gigante);
- possibilità di creare la propria voce sintetica personale (cioè una voce che abbia il proprio timbro).

Il motore di sintesi "Flexible and Adaptive Text To Speech" è un'applicazione web che espone le proprie funzionalità mediante interfaccia HTTP. Pertanto nello sviluppo si dovranno prevedere appositi meccanismi per consentire il servizio (utilizzando gli altri motori di sintesi presenti sul dispositivo), anche in assenza di connessione.

La progettazione e l'implementazione dovranno tener conto della riusabilità delle componenti sviluppate, per consentire la realizzazione di altre applicazioni, incluse quelle più tradizionali.

Pertanto si raccomanda la realizzazione di quattro componenti:

- modulo per la sintesi, seguendo le specifiche stabilite dalla piattaforma mobile di riferimento, cosicché il "Flexible and Adaptive Text to Speech" possa essere utilizzato da tutte le applicazioni che già utilizzano i servizi TTS di sistema;
- applicazione per la configurazione delle funzionalità aggiuntive;
- libreria per lo sfruttamento delle funzionalità aggiuntive;
- applicazione innovativa che utilizzi il motore di sintesi (direttamente o attraverso il modulo per la sintesi realizzato) e le funzionalità aggiuntive.

Al gruppo di lavoro verrà fornito da Mivoq supporto alla configurazione del motore di sintesi, nonché sull'utilizzo delle sue funzionalità. Verrà quindi fornito un insieme di voci realizzate con dati liberi, utilizzabili gratuitamente anche per scopi commerciali. Agli studenti del gruppo di lavoro verrà consentito, gratuitamente, di realizzare la loro voce sintetica, utilizzabile con questo motore di sintesi, attraverso un apposito servizio di Mivoq in fase di rilascio.

L'applicazione innovativa che verrà realizzata avrà come principale requisito che si sfruttino, in modo utile, le potenzialità della sintesi vocale non fornite dai servizi TTS di sistema. L'identificazione dell'applicazione specifica, la sua progettazione e la sua implementazione saranno responsabilità del gruppo di lavoro. Tale applicazione ha come scopo dichiarato quello dell'innovazione, cioè dell'introduzione di sistemi nuovi atti a modificare la prassi comune. È importante che il gruppo di lavoro capisca che difficilmente un processo di innovazione ha successo se non è possibile identificare un caso d'uso di facile comprensione, per cui la nuova proposta dovrà apportare dei vantaggi rispetto alle altre esistenti. Particolare attenzione, quindi, dovrà essere posta nella definizione del campo di utilizzo.

In figura 3 è possibile farsi un'idea degli strumenti a disposizione. I motori di sintesi vocale, generalmente, offrono

1. un metodo per la conversione da testo a voce, spesso chiamato "say()". Questo metodo accetta come parametri il testo da leggere e dei parametri, dipendenti dall'engine, per modificarne il risultato. Quasi sempre è possibile specificare la voce da utilizzare per realizzare l'audio e la localizzazione (espressa con un identificativo che indica sia la lingua che lo stato di riferimento);
2. vari metodi per interrogare il sistema a proposito dei valori ammissibili per le chiamate al metodo "say()". Questi metodi sono solitamente utilizzati per realizzare delle interfacce di configurazione, attraverso la creazione di appositi menu di selezione o altro.
    - "getLocalesList()": il metodo fornisce un elenco delle localizzazioni supportate ed è quasi sempre presente (FA-TTS include questa funzionalità);
    - "getVoiceList()": il metodo fornisce un elenco delle voci a disposizione ed è quasi sempre presente (FA-TTS include questa funzionalità);

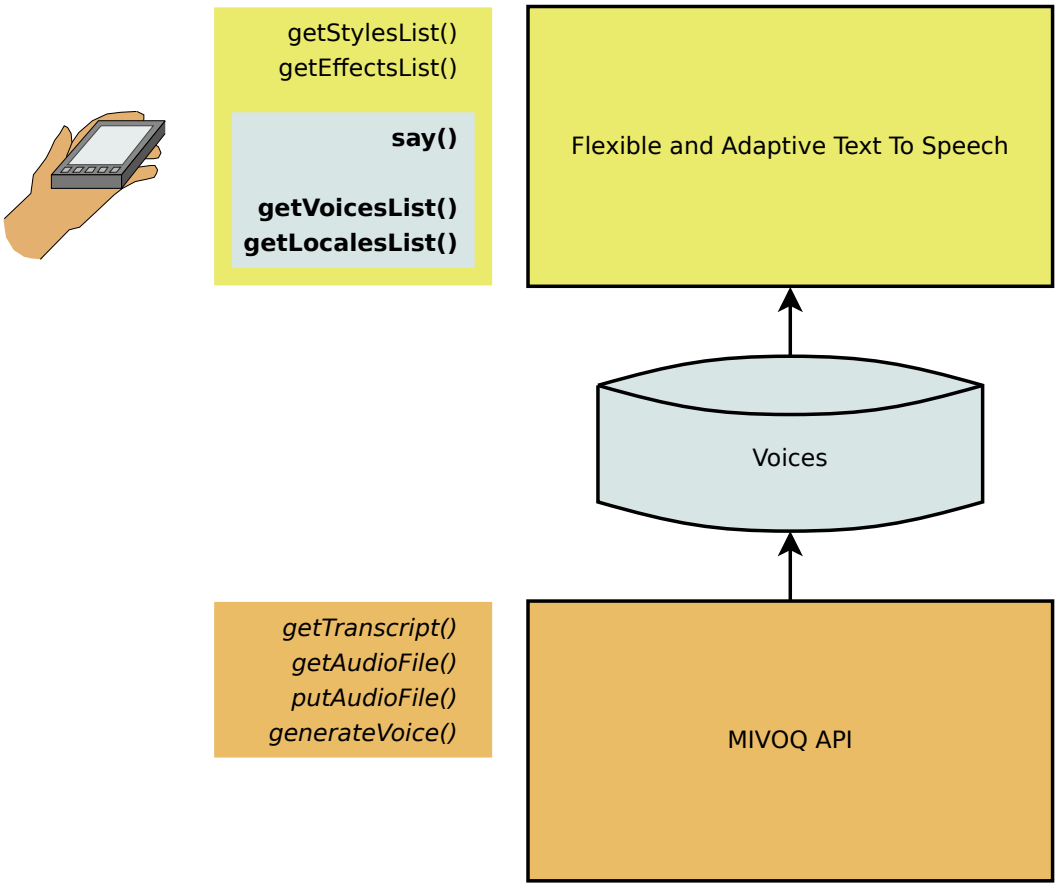

Figura 3: TTS API

- "getStylesList()": il metodo fornisce un elenco degli stili di lettura (es.: triste, felice, ...) a disposizione ed è raramente implementato (FA-TTS include questa funzionalità);
- "getEffectsList()": il metodo fornisce un elenco degli effetti che si possono applicare alla voce (es.: modifica dell'altezza della voce, effetto robotico, eco, ...) ed è raramente implementato (FA-TTS include questa funzionalità).

FA-TTS mette a disposizione tutte le funzionalità sopra elencate. Di contro le normali interfacce di configurazione presenti nei dispositivi mobili attuali permettono di selezionare unicamente la lingua (presentando all'utente le possibilità elencate da "getLocalesList()") e la voce (presentando all'utente le possibilità elencate da "getVoicesList()"), ma non permettono di configurare né lo stile, né eventuali effetti da applicare. Scopo del progetto è anche di colmare questa lacuna.

Mivoq offre un servizio di personalizzazione della voce (figura 4) attraverso cui è possibile permettere all'utente di creare la propria voce. Il servizio è disponibile sia attraverso una piattaforma web, che come API REST ed al momento è in fase di beta testing. L'utilizzo di un servizio sperimentale e non ancora testato può rappresentare un rischio per il progetto. Ciò non di meno Mivoq si impegna a fornire tutta l'assistenza necessaria ed a correggere eventuali errori nel più breve tempo possibile. Il servizio prevede che l'utente legga alcune frasi ad alta voce, le invii ai server Mivoq, che creeranno automaticamente una voce sintetica con il timbro vocale dell'utente stesso.

I metodi principali messi a disposizione dalla API REST sono:

- "getTranscript()": il metodo fornisce una frase da far leggere all'utente;
- "getAudioFile()": il metodo permette di ottenere il file audio relativo ad una frase letta;
- "putAudioFile()": il metodo permette di caricare un file audio relativo ad una frase da leggere;
- "generateVoice()": il metodo consente di avviare la creazione di una voce.

3 Specifiche di progetto e riferimenti

Considerata la natura di innovatività del progetto, viene lasciato un ampio margine in merito ai requisiti del sistema.

Per quanto riguarda le scelte tecnologiche si richiede:

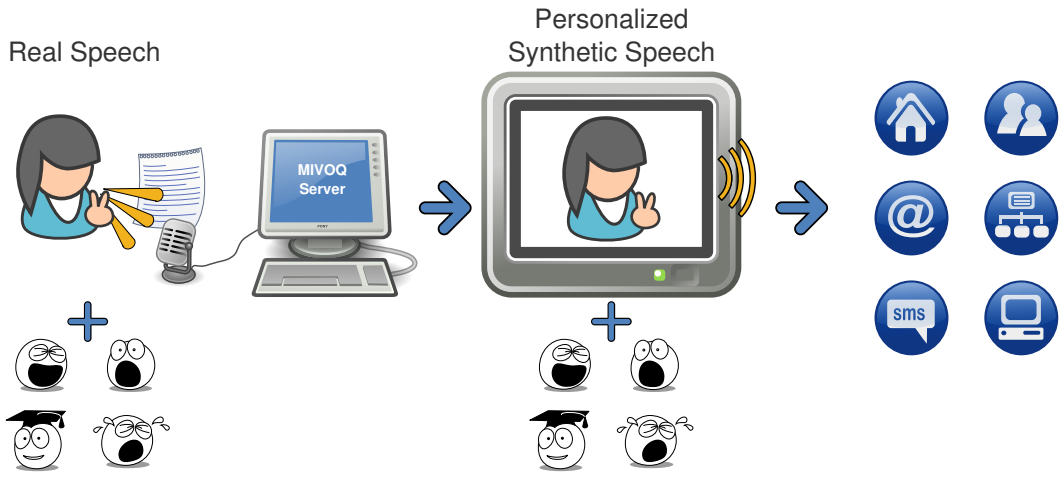

Figura 4: MIVOQ Personalizer
`http://www.mivoq.it`

- l'utilizzo del motore di sintesi "Flexible and Adaptive Text-To-Speech"
  `http://lab.mediafi.org/try-flexibleandadaptivetexttospeech.html`
  `http://lab.mediafi.org/discover-flexibleandadaptivetexttospeech.html`
- la realizzazione dell'applicazione per una o più piattaforme mobili a scelta fra:
  - Android `http://developer.android.com/index.html`
  - iOS `https://developer.apple.com/ios/download/`
  - Windows Phone `https://dev.windows.com/en-us/windows-apps`

Se lo si ritiene opportuno è possibile utilizzare anche un framework di sviluppo multipiattaforma come:

- Unity `https://unity3d.com/`
- PhoneGap `http://phonegap.com/`
- altro framework.

3.1 Obiettivi obbligatori

Gli obiettivi obbligatori del progetto sono:

- realizzazione di un'applicazione che utilizzi il motore di sintesi "Flexible and Adaptive Text-To-Speech" su almeno una piattaforma mobile
  - implementazione di meccanismi atti a risolvere le problematiche legate all'utilizzo di un servizio remoto
  - implementazione di un'interfaccia di configurazione dei servizi TTS realizzati grazie al motore di sintesi adeguata all'applicazione
- documentazione dell'applicazione
  - descrizione del caso d'uso
  - analisi dei requisiti
  - descrizione tecnica

3.2 Obiettivi desiderabili

- suddivisione del progetto in sotto componenti e loro implementazione separata
  - implementazione di un modulo per la sintesi che rispetti le specifiche per i servizi TTS della piattaforma;
  - implementazione di un'interfaccia di configurazione del modulo per la sintesi che permetta la configurazione anche di funzionalità aggiuntive. È importante cercare di mappare quanto più possibile le funzionalità che non sono rappresentabili secondo le normali specifiche della piattaforma, in funzionalità simili (per es.: se non è possibile utilizzare degli effetti, mappare gli effetti in voci aggiuntive...).
  - implementazione di un'API riutilizzabile per consentire l'accesso a tutte quelle funzionalità aggiuntive, normalmente non a disposizione dalla piattaforma;
  - implementazione dell'applicazione stessa.

3.3 Obiettivi opzionali

- supporto multipiattaforma;
- utilizzo e integrazione di servizi aggiuntivi, come ad esempio l'integrazione del servizio di personalizzazione della voce nell'applicazione o l'utilizzo di risorse esterne per ottenere contenuti.

4 Possibili idee

In questa sezione sono presenti alcune applicazioni che la proponente ritiene sufficientemente innovative. Il fornitore è libero di realizzarle, di prendere spunto o di ignorarle.

4.1 Applicazione per la realizzazione di piccoli sceneggiati

L'applicazione sfrutta la possibilità di avere vari stili e varie voci a disposizione per permettere ai suoi utilizzatori di raccontare piccole storie attraverso la sintesi vocale. L'applicazione dovrebbe rendere agevole la creazione e la modifica di questi sceneggiati, la loro composizione in un file video e la condivisione su social network. Una buona idea potrebbe essere anche consentire l'integrazione di suoni (musiche di sottofondo, rumori, ...) e immagini, magari sfruttando le numerose risorse gratuite (e non) disponibili online.

Applicazioni simili sono state realizzate per personal computer, come ScriptVox di ScreamingBee (fig. 5), che consente di creare dei personaggi, assegnando una voce ad ognuno e aggiungere ad associare ad ogni sezione di testo un personaggio specifico.

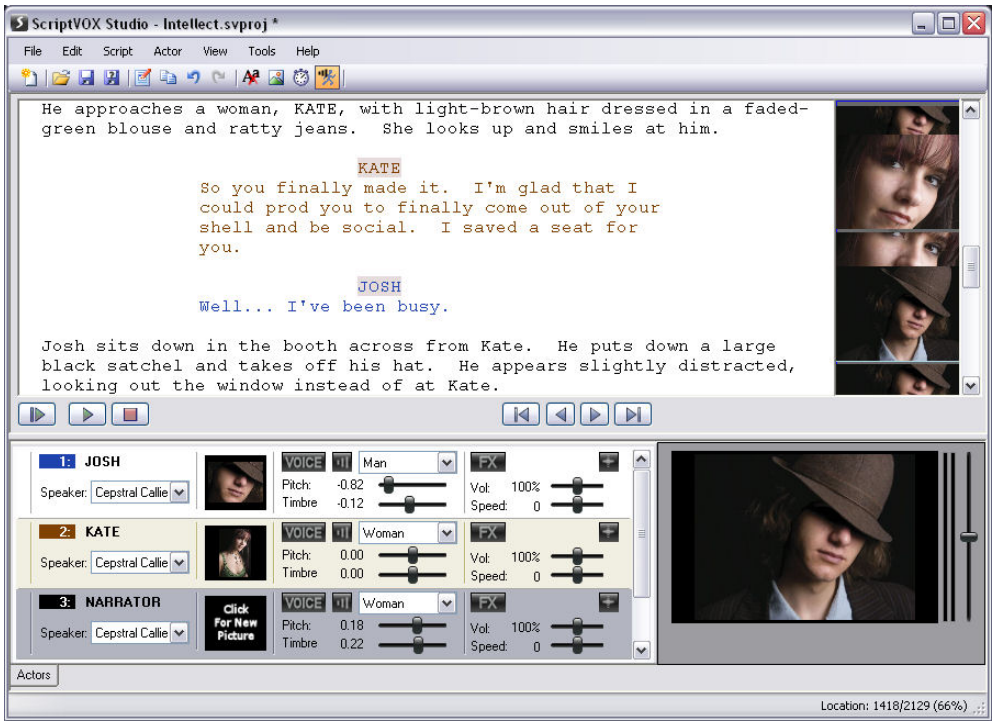

Figura 5: ScriptVox
`http://www.screamingbee.com/product/ScriptVOXStudio.aspx`

4.2 Applicazione per permettere di parlare a chi non può usare la voce

Applicazioni di questo genere esistono, già, come Free Speech (fig. 6).

Quello che si può aggiungere in questo caso è la possibilità di applicare effetti alla voce sintetica per replicare emozioni (contentezza, rabbia, ...) o realizzare effetti divertenti. L'applicazione potrebbe avere come utenti finali sia le persone che non possono più utilizzare la loro voce, sia quelli che vogliono temporaneamente cambiare la loro voce con una virtuale con obiettivo ludico.

Applicazioni molto simili concettualmente, esistono anche con obiettivi diversi, come per esempio Avaz Free Speech (fig. 7).

In questo caso l'obiettivo è di poter aiutare i bambini all'apprendimento della lingua permettendo loro di comporre frasi utilizzando immagini anziché parole vere e proprie. In questo caso il valore aggiunto potrebbe essere di poter utilizzare una voce che più si adatta all'utente finale (e.g.: la voce della mamma, o la voce di un

Figura 6: Free Speech
`https://play.google.com/store/apps/details?id=com.blogspot.tonyatkins.freespeech`

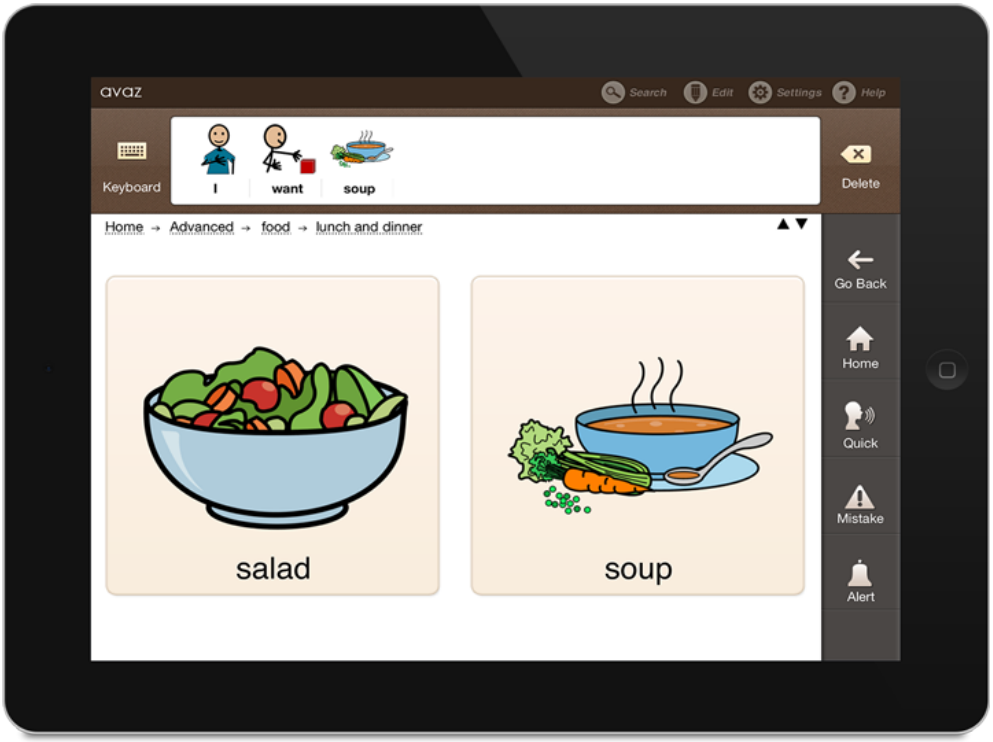

Figura 7: Avaz Free Speech
`http://avazapp.com/freespeech/`

bambino). Un'altra possibilità potrebbe essere di realizzare la frase in una lingua non nota a priori. In questo caso poter utilizzare la propria voce o poter applicare degli effetti per trasmettere emozioni, potrebbe essere utile per dialogare con persone che parlano lingue diverse dalla propria.

4.3 Applicazione per l'arricchimento dei messaggi per il navigatore

In questo caso l'applicazione riceverà i messaggi del navigatore e dovrà, sulla base del loro contenuto, modificarli o arricchirli, simulando una personalità propria, selezionando gli effetti con cui modulare la propria voce di conseguenza.

Nel mercato sono già state proposte applicazioni similari, come per esempio le voci personalizzate per il navigatore GPS TomTom, fra cui quella di Marco Ranzani (fig. 8). In questo caso, però, l'applicazione avrebbe carattere

più generale e si potrebbe adattare a più contesti. Inoltre l'aggiunta di nuove frasi, non richiederebbe l'intervento ulteriore dello speaker.

Figura 8: TomTom - Ranzani

4.4 Applicazione per la notifica di eventi

L'applicazione sfrutta la possibilità di caratterizzare le notifiche secondo vari stili, a seconda della loro tipologia (per es.: utilizzo di un tono spaventato per segnalare la batteria scarica). Una buona idea potrebbe essere anche di permettere la configurazione degli stili associati ad ogni messaggio di notifica e alle notifiche ripetute (per poter permettere, ad esempio, l'utilizzo di una voce annoiata all'ennesimo messaggio in chat non letto).

A Note sul capitolato d'appalto

Questo capitolato d'appalto rappresenta una prima bozza di lavoro per aiutare il fornitore a capire le tematiche del progetto e del tipo di sviluppo che sarà necessario.

A.1 Aspettative della proponente

La proponente si aspetta di ottenere un'applicativo che possa essere utilizzato per dimostrare efficacemente le potenzialità del motore di sintesi "Flexible and Adaptive Text-To-Speech", senza prerogative di ulteriore sfruttamento dal punto di vista commerciale. L'applicazione realizzata, salvo ulteriori accordi, resterà di proprietà esclusiva del fornitore.

La proponente potrebbe avere interesse a riutilizzare parte del codice sorgente prodotto ed incoraggia, ove possibile, il rilascio con licenze opensource in stile BSD/MIT per le componenti di carattere generale ed in particolare per eventuali moduli di sintesi, librerie e interfacce di configurazione.

A.2 Contatti

La proponente potrà essere contattata in qualsiasi momento attraverso l'indirizzo email `tech@mivoq.it`, al quale risponde il reparto tecnologico dell'azienda. Le email dovranno contenere in oggetto la sigla "[UNIPD-TTS]" e dovranno essere indirizzate all'attenzione di Giulio Paci, che sarà il referente principale per il progetto.

In alternativa è possibile telefonare al numero 0490998335, al quale risponde il reparto tecnologico dell'azienda. In questo caso sarà sufficiente specificare che si chiama a proposito del progetto di "Ingegneria del software".

B Note sulla proponente

Mivoq è una startup, nata nel 2013, che si occupa di sintesi vocale, con lo scopo di permettere ad ogni singolo utente di avere la propria voce digitale che possa rappresentarlo anche dove normalmente non è presente, attraverso applicazioni innovative per la lettura di SMS con la voce del mittente, la lettura di post su Facebook con la voce dell'autore, la personalizzazione di assistenti virtuali con la voce dell'utente o la produzione della propria voce quando questa è venuta meno a causa di un'operazione o di una malattia (si pensi per esempio al fisico Stephen Hawking).

L'obiettivo di Mivoq non è di realizzare queste applicazioni, ma di fornire la tecnologia necessaria a fare in modo che vengano realizzate. Con questo progetto intendiamo mettere alla prova questa tecnologia ed ottenere un esempio concreto della sua efficacia.