SVEUČILIŠTE U ZAGREBU
Fakultet elektrotehnike i računarstva

VOICE SYNTHESIS / SINTEZA GOVORA

Seminarski rad iz kolegija
PODATKOVNI VIŠEMEDIJSKI PRIJENOS I RAČUNALNE MREŽE

Studenti:

Violeta Đurđek
Karlo Ercegović
Barbara Šmit

ŠTO JE SINTEZA GOVORA

Sinteza govora predstavlja operaciju pretvaranja pisanog ulaza u govorni izlaz. Ulaz može biti u obliku grafemske, ortografske ili fonemske skripte, ovisno o izvoru. Jednostavnije rečeno, sinteza govora je umjetno generiranje ljudskog govora.
Sustavi koji se za to koriste nazivaju se sintetizatori govora, a mogu biti implementirani kao softver ili hardver.

Sinteza govora često se kraće naziva Text-to-Speech (TTS), obzirom da upravo i pretvaraju tekst u govor.

Postoji nekoliko algoritama za sintezu govora. Izbor algoritma ovisi o operaciji koju želimo izvršiti. Najjednostavniji način je jednostavno snimiti glas osobe koja govori željene izraze, ali to predstavlja samo ograničen izvor fraza i rečenica. Kvaliteta ovisi o načinu snimanja.

Sofisticiraniji, ali lošije kvalitete su algoritmi koji dijele govor u manje jedinice. Najčešće korištena jedinica je fonem, najmanja lingvistička jedinica. Ovisno o jeziku, postoji oko 35-50 fonema u zapadno-europskim jezicima. Problem je u kombiniranju fonema jer tečan govor zahtjeva tečan prijelaz između elemenata (fonemskih jedinica). Razumljivost je stoga manja, no mala je i zahtjevnost na memoriju.

Rješenje ovog problema je korištenje difona. Umjesto dijeljenja u prijelazima, stanka se radi u sredini fonema, što ostavlja prijelaze netaknute. To daje oko 400 elemenata i kvaliteta raste.

Što su dulje te jedinice, postoji više elemenata, ali uz potrebnu memoriju raste i kvaliteta. Ostale jedinice koje su u širokoj primjeni koriste su poluslogovi, slogovi, riječi ili njihova kombinacija.

Načini sinteze govora

Postoje dva glavna načina za generirnje valnih oblika umjetnog govora:

- lančana sinteza (*engl.* Concatenative synthesis)
- formant sinteza (*engl*. Formant synthesis)

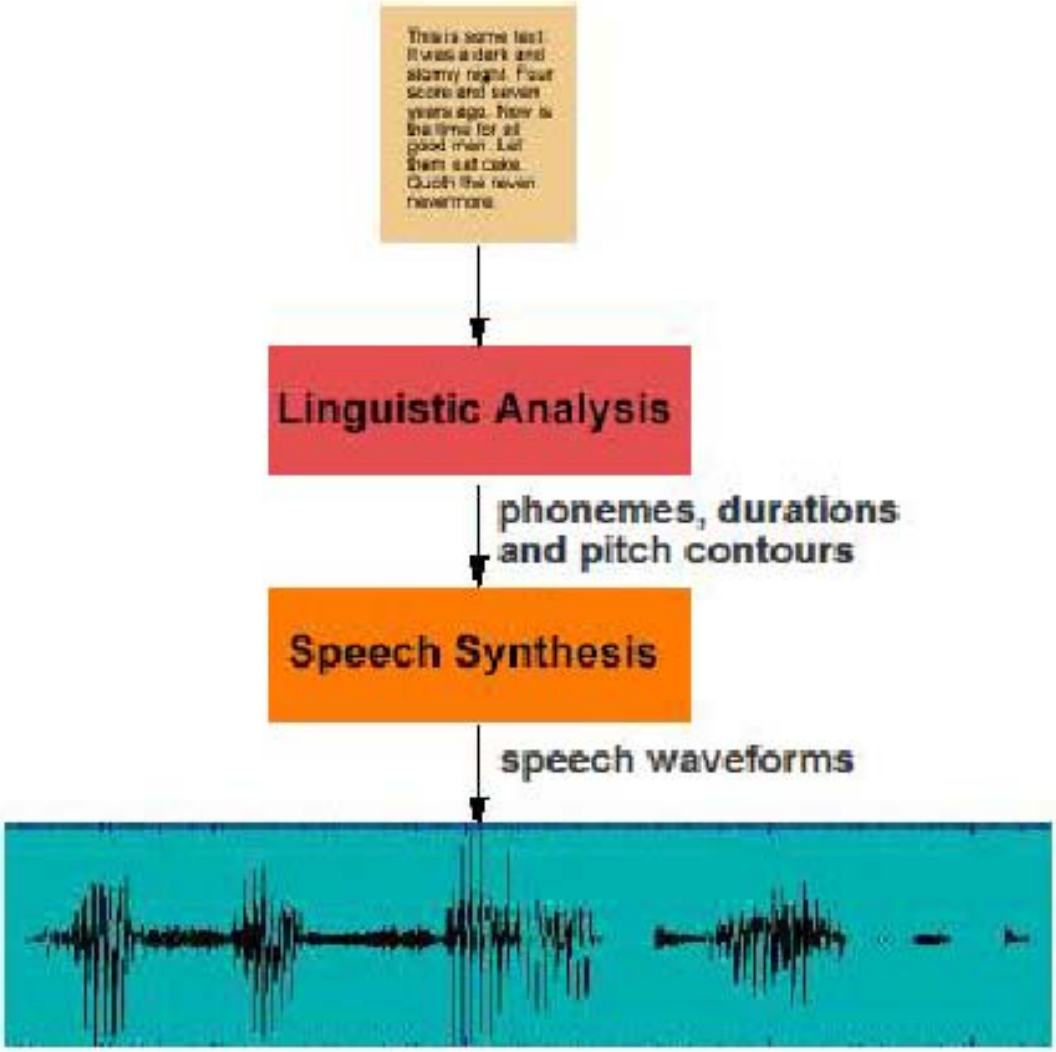
Lančana sinteza se bazira na spajanju (ili nizanju) segmenata snimljenog govora. Općenito, lančana sinteza generira umjetni glas najsličniji prirodnom ljudskom govoru. Ipak, prirodne varijacije u govoru i automatizirane tehnike za segmentaciju valnih oblika ponekad rezultiraju zastajkivanjem izlaza umanjujući prirodnost glasa. Tri su glavna podtipa lančane sinteze: *jedinično selektivna sinteza, difona sinteza* i *područno specifična domena.*

Formant sinteza ne koristi uzorke ljudskog glasa, već umjetni glas kreira korištenjem akustičnog modela. Parametri kao što su temeljna frekvencija, zvučnost i razina šuma su varirani tokom vremena da bi stvorili valni oblik umjetnog govora.
Mnogi sustavi temeljeni na formant sintezi generiraju umjetni robotski glas i izlaz nikad ne bi mogao biti zamijenjen s glasom pravog čovjeka. Maksimalna prirodnost nije uvijek cilj tih sustava, stoga formant sinteza ima nekih prednosti u usporedbi s lančanom metodom.

Ostali načini sinteze (manje korišteni):

- artikulacijska sinteza
- hibridna sinteza
- HMM-bazirana sinteza (Hidden Markov Model)

Komponente TTS sustava:



PREGLED ALATA ZA SINTEZU GOVORA

MBrola

MBROLA je visoko-kvalitetni, difono bazirani sintetizator govora, besplatno dostupan. Ostvaren je od TCTS laboratorija fakulteta *Faculte Polytechnique de Mons* (Belgija) s ciljem da osigura set govornih sintetizatora za što je moguće više jezika i svima dostupnima za ne-komercijalnu primjenu.

Temelj MBrola projekta je *MBROLA govorni sintetizator* baziran na spajanju difona. Uzima listu fonema kao ulaz zajedno s prozodičnim informacijama (trajanje fonema i po dijelovima linearan opis pitch-a) i generira 16-bitne (linearne) uzorke govora na frekvenciji uzorkovanja korištene baze podataka difona → stoga MBrola zapravo NIJE govorni sintetizator jer ne može raditi sa ulazom u obliku pisanog teksta. Baza podataka difona pripojena MBrola formatu potrebna je za rad sintetizatora.

Moguće je poslati vlastitu snimku govora koja će biti spremljena u Mbrola bazu podataka za sintezu. Trenutno postoji baza podataka za sljedeće jezike: američki engleski, brazilski, portugalski, bretonski, britanski engleski, francuski, njemački, grčki, rumunjski, španjolski i švedski.

TEXT-TO-SPEECH SUSTAVI

Festival

Festival je najkompletniji besplatan sustav za sintezu uz opsežan priručnik. U cjelini nudi kompletnu pretvorbu teksta u govor uz različite API-e (**a**pplication **p**rogramming **i**nterface), kao i okruženje za razvoj i istraživanje tehnika sinteze govora.
Sustav je napisan u C++-u s komandnim interpretatorom za generalnu kontrolu baziranom na Scheme programskom jeziku. Višejezičan je, trenutno podržava engleski (američki i britanski) i španjolski.
Na *home page* stranici mogu se pronaći demo snimke, kompletni priručnik i pristup download stranici. Uključuje kompletan izvor i dokumentaciju (FSF texinfo format), leksikone i govornu bazu podataka za pretvorbu (britanskog) engleskog teksta u govor.

KARAKTERISTIKE:

- engleski (britanski i američki) i španjolski text-to-speech
- eksterno podesivi jezično-neovisni moduli: leksikoni, pravila pretvorbe slova u glas, token-izacija, intonacija i trajanje, tekst modovi, odabir tipa sinteze difon/jedinica
- waveform sintetizatori
    - difono bazirani
    - MBrola podrška
- prenosiva distribucija, on-line dokumentacija

WinSpeech

WinSpeech je text-u-govor aplikacija koja čita tekst i producira govor na audio izlazu. Ima osnovne alate za uređenje teksta, omogućen je govor iz trenutno napisanog rada, podržava DDE server koji omogućava drugim Windows aplikacijama da šalju tekst za izgovor, ima mod rada za učenje koji pruža audio instrukcije za vrijeme rada programa, te alate za uređenje rječnika za proizvoljni odabir izgovora.
WinSpeech je *shareware* program proizveden od PCWholeWare.
WSPLIB tekst-u-govor DLL je biblioteka govornih funkcija za razvoj.

KARAKTERISTIKE:

- *file* mod – čitanje ASCII teksta ili Windows WRITE (.WRI) datoteka
- *clipboard* mod- rad s drugim Windows aplikacijama kao dodana govorna usluga
- *DDE server* mod – radi kao govorni alat, dozvoljava ostalim aplikacijama da iskažu tekst preko DDE kanala za govor
- *command line* mod –čita nizove podataka ili datoteke specificirane u parametrima komandne linije za vrijeme podizanja (start up) programa
- alat za uređivanje osnovnog teksta → učitaj, spremi, kopiraj, zalijepi i editiraj tekst
- audio vodič s instrukcijama za vrijeme rada programa
- podesivi pitch i brzina izgovora
- alat za uređivanje rječnika za prilagodbu izgovora
- driver za interni govorni uređaj može se instalirati za vrijeme instalacije zvučne kartice PC-a

BaBel Technologies

BaBel Technologies nudi vrhunsku sintezu govora zahvaljujući Multi Band Resynthesis OverLap Add tehnici (MBrola). Nova tehnologija sinteze je patentirana 1996 g. i nagrađena je s European Information Technology Prize iste godine za svoj inovativni pristup tzv. spajanoj sintezi govora.

Visoko kvalitetna sinteza govora uz malu zahtjevnost CPU-a: nova generacija visoko-kvalitetnih sintetizatora govora više ne ostavlja dojam slušaocu da je zvučnik stroj. Riječi su glatko *izgovarane* vremenski baziranim difonim spajajućim algoritmom. Ovaj algoritam je prvi koji dozvoljava izglađivanje spektra uz zadržavanje vrlo niskih računalnih troškova. Nije potreban DSP. Standardni Pentium 100 može pokrenuti sintetizator 20 puta brže prema realnom vremenu.

Višejezična sinteza govora: MBrola tehnika koristi baze podataka ovisne o jeziku i govorniku da bi proizvela bilo koju rečenicu na danom jeziku i s datim glasom. MBrola govorni sintetizator je dostupan za engleski, njemački, francuski, nizozemski, brazilski portugalski, španjolski, švedski i rumunjski jezik. Ostali jezici su trenutno u izradi.

Više jezika i glasova na zahtjev: ova usluga uključuje razvoj skrojenih glasova i jezika. Difona baza podataka tipične je veličine do 5Mb i, uz korištenje prednosti MBrola formata, ovo specifično kodiranje kompresira baze podataka u omjeru 7:1.

Infovox Desktop

Infovox Desktop je difono bazirani **BaBel Technologies** TTS alat razvijen uz pomoć biblioteke unaprijed snimljenih ljudskih izgovorenih difona za što prirodniji sintetizirani glas. Nudi neograničenu primjenu aplikacijskih mogućnosti za razvojne programere softvera koji žele koristiti sintetički govor kao nositelj informacije, i za one koji žele integrirati sintetički govor u svoje proizvode ili usluge.
Trenutno je dostupan za: britanski engleski, finski, islandski, španjolski, danski, francuski, talijanski, nizozemski, švedski, njemački i norveški jezik.

KARAKTERISTIKE:

- difono bazirani TTS softver, uz mogućnost formant sinteze
- korisnički leksikon i podrška
- poboljšani Voice Manager (upravljač glasom)
- „Key Speaker“ funkcija → slovka i čita bilo koji Window program
- grafički ekvalizator
- varijabilna brzina čitanja

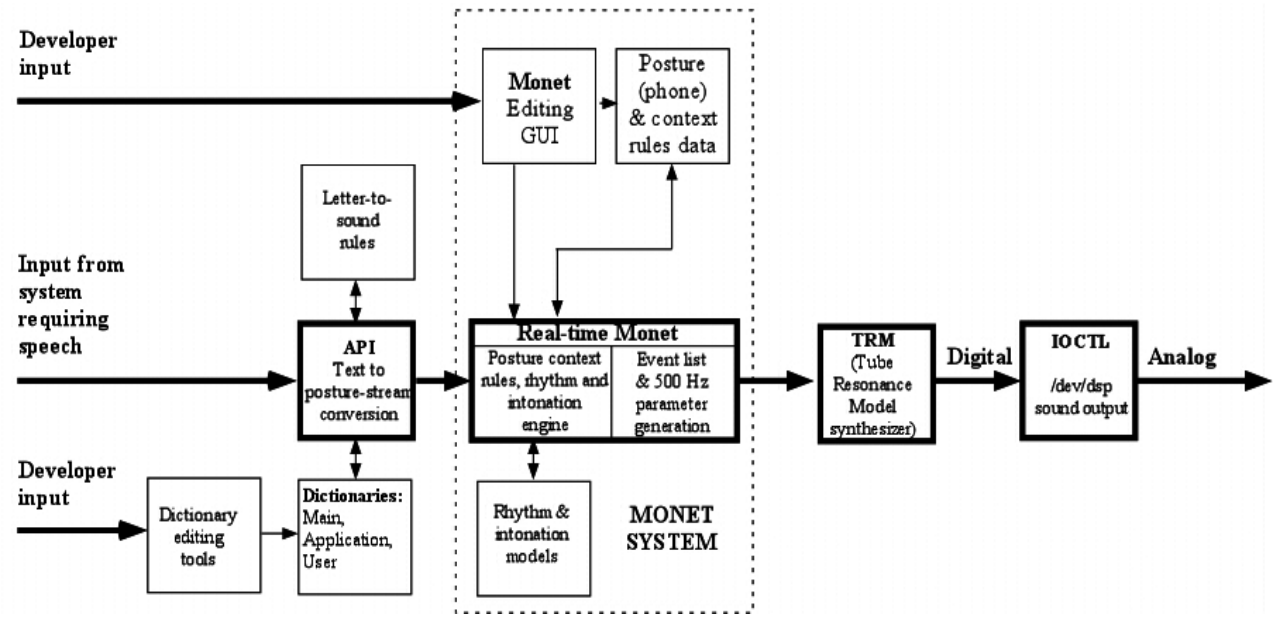
Gnuspeech

Gnuspeech je prilagodljiv TTS paket, baziran na real-time, artikularnoj, upravljivoj sintezi govora. Konvertira tekstovne nizove podataka u foneme podržane rječnikom za izgovor, po pravilima pretvorbe slova u zvuk, ritma i intonacije. Transformira foneme u parametre za nisku razinu pretvorbe artikulacijskog sintetizatora i zatim proizvodi artikulatorni model ljudskog glasa u obliku izlaza pogodnog za standardne audio izlaze. Gnuspeech se još uvijek razvija.

KARAKTERISTIKE:

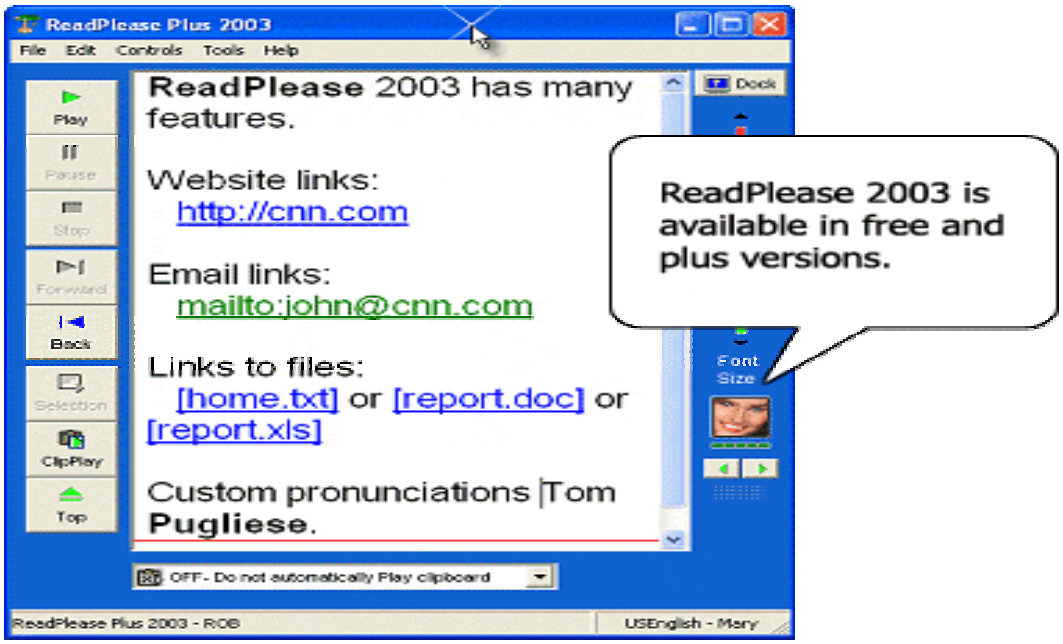
- TRM model (Tube Resonance Model = rezonantni cijevni model) ljudskog vokalnog trakta (poznat i kao waveguide model) koji vrlo dobro imitira fizičke karakteristike trakta
- kontrolni model za TRM baziran na analizi osjetljivosti formanta koji omogućava preciznu specifikaciju relevantne konfiguracije vokalnog trakta za govor i uključuje nisko razinski artikulatorni model uz mali broj parametara i niski bit rate
- baze podataka specificiraju artikulaciju i kontrolu dinamike koje su potrebne za produkciju engleskog jezika iz povećanog fonemskog ulaza; neki francuski vokali su također uključeni
- *«Monet»* - GUI-bazirani sustav za kreiranje i uređivanje baze podataka omogućava upravljanje i mijenjanje fonemičkih podataka i dinamičkim pravilima
- 70000+ riječi u rječniku engleskog izgovora s pravilima deklinacije za množinu i priloge; rječnik također sadrži informacije o govornim cjelinama za gramatičku analizu i uključuje 6000 osobnih imena
- pravila pretvorbe slova u zvuk koja se bave slovkanjem i riječima koje nisu u rječniku
- alati za upravljane rječnikom i iznošenje analize govora
- *«Synthesizer»* - GUI-bazirana aplikacija koja dozvoljava eksperimentiranje sa samostalnim TRM-om; svi parametri su varijabilni, a izlaz je nadziran i analiziran

Na slici je shematski prikaz strukture GNUSpeech sustava:



ReadPlease 2003 i ReadPlease Plus 2003

TTS softver za Windows bazirane operacijske sustave. Čita bilo koji tekst prikazan na ekranu – višenamjenski alat jednostavan za uporabu.



KARAKTERISTIKE ReadPlease 2003:

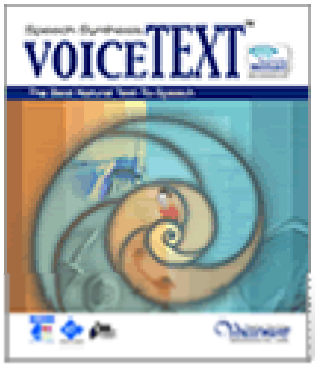
- Microsoft glasovi: Mike, Mary, Sam
- podesivi font i boja pozadine
- čitanje teksta preko Windows međuspremnika (clipboard) iz bilo kojeg programa (copy/paste)
- kontrolno čitanje iz programske trake sustava (system tray)
- čitanje e-mail emoticona kao što su ☺ i ☹
- podesiva brzina izgovora
- *ReadPlease Enable* za web stranice

DODACI ReadPlease Plus 2003 verzije:

- brzi forward i backward
- označavanje (highlight) teksta za vrijeme čitanja
- *Dock Mode* – smjestiti će se na vrhu ekrana
- dodavanje vlastitih riječi i izgovora
- PLAY može započeti bilo gdje u dokumentu
- Podesiva pauza između paragrafa
- *Hot-key* kontrolne tipke za sve funkcije
- *AT&T Natural Voices* kompatibilno
- višejezičnost

VoiceText

VoiceText je vodeći softver za sintetiziranje umjetnog glasa iz teksta. Dostupan je u konfiguraciji za širok raspon ugrađenih uređaja, desktop i mrežnih/serverskih aplikacija, što ga čini vrlo fleksibilnim visoko-kvalitetnim TTS rješenjem na današnjem tržištu.
Dostupan je na US engleskom, korejskom, japanskom i mandarin kineskom.



KARAKTERISTIKE:

- prirodno zvučeći glas i čist izgovor, izlaz je dobro razumljiv
- konfigurabilan je za upotrebu u vrlo širokom rasponu ugrađenih, desktop i mrežnih/serverskih aplikacija, velika fleksibilnost primjene
- višejezičnost – američki engleski, korejski, japanski i mandarin kineski, kroz jezike je dostupna kolekcija od 11 izvornih (materinskih) jezika
- velik, proširiv rječnik – tisuće izgovora su uključeni u standardni rječnik svakog podržanog jezika, omogućeno je podešenje rječnika stoga razvojni programeri mogu prilagoditi izgovor simbol, kratica i novo unesenih izraza
- ekspresivna kontrola – pitch, brzina, glasnoća i pauze su podesive dinamički i kao standardne postavke
- predobrada ulaznog teksta –VoiceText automatski obrađuje specijalne ulaze kao što su datumi, vrijeme, kratice u adresama, te rečenice na kombiniranim jezicima, nove grupe mogu biti dodane korištenjem podesivih pravila
- dizajn sustava – VoiceText sintetizira govor u sub-realnom vremenu i podržava višenizne i višekanalne arhitekture; opcionalni djelitelj je dostupan za mrežne/serverske konfiguracije
- fleksibilan format izlaza, uključujući 8kHz/16kHz uzorkovane, linearno 8-bit/16-bit PCM, 8-bit mu-law/a-law, ADPCM, Windows .wav i dr.
- podrška za API-e (Application Programming Interface) – podržava SAPI4, SAPI5, C/C++, COM i Java-bazirane API-e

Loquendo Text-to-Speech (TTS)

Loquendo TTS softvare sintetizira glas vrlo blizak prirodnom za dinamičke podatke, te za serverski bazirane, multimedijalne, PDA, ugrađene i multimodalne glasovne aplikacije. Loquendo „*Unit Selection*“ lančana tehnika primjenjiva je na vrlo širok raspon glasovnih uzoraka pomoću kojih se mogu stvoriti novi visoko-kvalitetni glasovi. Osigurava Loquendovo tržišno vodstvo u kvaliteti, učinkovitosti, prenosivosti, prirodnoj boji glasa i intonaciji, te točnosti izgovora. Omogućava: čitanje e-maila, real-time vijesti, pristup korporacijskim dokumentima, automotivnu telematiku (informacijska i komunikacijska tehnologija), primjenu na bilo koju ugrađenu aplikaciju

Loquendo TTS baza ekspresivnih glasova i osobnosti iz cijelog svijeta je stalno rastuća. Efikasne razvojne metode garantiraju brzu ponudu novih visoko-kvalitetnih glasova i jezika. Loquendo također podržava izgradnju specifičnih glasova koji se podudaraju sa određenim pojedincem.

Loquendo glasovi su čisti, prirodni i tečni i obogaćeni su sa repertoarom tzv. ekspresivnih znakova: pozdravi i eksklamacije, interjekcije i paralingvistički događaji, koji sugeriraju ekspresivnu namjeru (potvrda, oklijevanje, zahvala, itd.).

Loquendo TTS algoritam je vrlo efikasan, pa su zahtjevi na procesor minimizirani i garantiran je iznimno brzi odgovor. Može suvremeno sintetizirati različite jezike i glasove, te po želji prelaziti s jednog jezika na drugi i za vrijeme rada (*Voice switching*).
Loquendo TTS *Director* daje razvojnim programerima potpunu kontrolu nad vlastitim glasovnim aplikacijama da mogu ekstenzivno poboljšati glasovne mogućnosti.
Mixed Language Capability omogućava ispravan izgovor stranih riječi bez potrebe za promjenom trenutnog glasa i jezika. *Audio Mixer* omogućava potpunu kontrolu audio izvora (glazba ili zvuk, različito samplani i/ili kodirani) → intermiksanje, sinkronizaciju ili repeticiju sa umjetnim (sintetiziranim) govorom. *Expressive cues* omogućava TTS korisnicima da proizvedu glas vrlo približan vlastitom glasu. *Pronunciation lexicon* osigurava da bilo koji specijalni vokabularni izraz, kratice, akronimi, čak i razlike u izgovoru narječja, zvuči upravo onako kako je developer to zamislio. Karakteristike svakog glasa (npr. pitch, brzina izgovora, glasnoća) mogu se vrlo fino regulirati i kontrolirati. Specijalni formati kao što su telefonski brojevi, valute i e-mail zaglavlja su ispravno izgovoreni. Ima ugrađeni detektor jezika koji automatski prepoznaje jezik svakog teksta. Dostupan je za: talijanski, španjolski, francuski, njemački, brazilski, portugalski, mandarin kineski, nizozemski, britanski i američki engleski, grčki, meksički, čileanski, američki španjolski, argentinski, švedski i katalanski jezik.

Ukratko → s Loquendo TTS jednostavno je dobiti sintetizirani glas upravo kako ga želite!

KARAKTERISTIKE:

- leksikon izgovora – korisniku moguće definirati izgovor (akronimi, strani nazivi, itd.)
- Mixed Language Capability
- Audio Mixer
- dinamička izmjena između više različitih glasova
- e-mail predobrada
- fleksibilna glasovna kontrola – kreiranje specijalnih efekata, mijenjanje brzine izgovora i pitch-a
- prilagodljivi glasovi
- prilagodljivi leksikon i baza podataka za Automotive
- ekspresivni TTS

TextAloud

TextAloud je TTS softver koji koristi sintezu govora za pretvorbu ulaznog tekstovnog dokumenta u govor u audio formatu za upotrebu u različite svrhe, npr. slušanje na PC-u, pretvorba u MP3 ili WMA datoteke koje se mogu koristiti na prijenosnim uređajima kao što su iPod, PocketPC ili CD playeri. Omogućit će produktivniji rad na kompjuteru (čitanje e-maila, web stranica, izvješća i dr.) ili jednostavno služiti za zabavu → TextAloud je praktičan i jednostavan čitač teksta.

KARAKTERISTIKE:

- čitanje e-maila, web stranica, izvješća,...
- pretvorba teksta u govor, MP3 ili WMA datoteke
- opcionalan izbor rada s AT&T Natural Voices, NeoSpeech Voices ili Cepstral Voices
- *Multi-Article* mod rada – neograničen broj otvorenih dokumenata
- podesivi trenutno odabrani glas, čak i za vrijeme govora, podesivi pitch
- podrška za tekst dokumente (.txt), rich text dokumente (.rtf), Microsoft Word dokumente (.doc), lokalne HTML dokumente i lokalne PDF dokumente (.pdf)
- *Batch File Converter* – omogućava pretvorbu većeg broja tekstovnih dokumenata u audio format
- jednostavna korekcija teksta pomoću *Proofread HotKey*
- Internet Explorer *Plug-In* omogućava jednostavnije čitanje teksta na webu
- podesiva frekvencija uzorkovanja (*sample rate*) pri stvaranju wave dokumenata
- jednostavan je za uporabu

TextToSpeech Kit

TextToSpeech Kit radi neograničenu pretvorbu engleskog teksta u sintetizirani govor u stvarnom vremenu. Dolazi u 2 paketa: *Developer kit* (razvojni alat) i *User Kit* (korisnički alat). *Developer Kit* omogućava stvaranje test aplikacija koje sadrže TTS. *User Kit* je podskup *Developer Kit-a*, i također podržava aplikacije koje sadrže TTS.

KARAKTERISTIKE:

- kontrola brzine govora, median pitcha, stereo balansa, glasnoće i intonacijskog tipa
- može biti izgovoren tekst bilo koje duljine, opcionalno poruka može biti na čekanju kod višestrukih aplikacija
- real-time kontrole kao npr. pauza, nastavak i brisanje
- izgovor se izvodi prema glavnom rječniku koji ima otprilike 100 000 ručno uređenih izgovora
- uključuje TTS server, TTS objekte, editore za izgovaranje, nekoliko primjera, fonetička slova, primjere kodova i dokumentaciju proizvođača

Tablični prikaz pregledanih alata za sintezu govora

ALAT	TEŽINA KORIŠTENJA	JEZICI	CIJENA
MBrola	srednje	višejezičan	freeware
Festival	teško	engleski, španjolski	freeware
WinSpeech	jednostavno	engleski	48$
Infovox	srednje	višejezičan	50$
GnuSpeech	srednje	engleski	freeware
ReadPlease 2003	jednostavno	engleski	freeware
ReadPlease Plus 2003	jednostavno	višejezičan	60$
VoiceText	jednostavno	višejezičan	25$
Loquendo	teško	višejezičan	150€ (1 voice, 1 lang.)
TextAloud	jednostavno	višejezičan	30$
TTS Kit	srednje	engleski	*User* 150$ *Developer* 250$

LITERATURA:

http://www.disc2.dk/tools/SGsurvey.html

http://linux-sound.org/speech.html

http://www.speechandhearing.net/laboratory/tools.html#synth

http://www.drspeech.com/VoiceSynthesis.html

http://www.bluechillies.com/list.html?k=voice+synthesis

http://encyclopedia.thefreedictionary.com/Voice+synthesis