Síntesi de la parla

Helenca Duxans Barrobés
Marta Ruiz Costa-jussà

PID_00188055

UOC
Universitat Oberta de Catalunya

Els textos i imatges publicats en aquesta obra estan subjectes –llevat que s'indiqui el contrari– a una llicència de Reconeixement-NoComercial-SenseObraDerivada (BY-NC-ND) v.3.0 Espanya de Creative Commons. Podeu copiar-los, distribuir-los i transmetre'ls públicament sempre que en citeu l'autor i la font (FUOC. Fundació per a la Universitat Oberta de Catalunya), no en feu un ús comercial i no en feu obra derivada. La llicència completa es pot consultar a http://creativecommons.org/licenses/by-nc-nd/3.0/es/legalcode.ca

Índex

Introducció 5

Objectius 6

1. Introducció a la síntesi de la parla 7

2. Aplicacions dels convertidors de text a veu 9

3. Els convertidors de text a veu 10
3.1. El mòdul de processament de llenguatge natural 10
3.1.1. Analitzador morfosintàctic 11
3.1.2. Transcriptor fonètic 11
3.1.3. Generador prosòdic 12
3.2. El mòdul de processament del senyal 13

4. Síntesi per concatenació 14
4.1. Elements dels sistemes de síntesi per concatenació 14
4.2. Unitats acústiques i construcció del corpus 15
4.3. Selecció de les unitats 15
4.4. Concatenació 16
4.4.1. Concatenació en el domini temporal 16
4.4.2. Concatenació en el domini paramètric 17

5. Modificacions prosòdiques 19
5.1. Tècniques basades en TD-PSOLA 19
5.1.1. Modificacions d'altura tonal 19
5.1.2. Modificacions de duració 21
5.1.3. Alternatives a l'algoritme de TD-PSOLA 23

6. Mesures de qualitat de la veu sintetitzada 24

Introducció

En aquest mòdul introduïm el concepte de síntesi de la parla i les aplicacions que té. En l'estat actual de la tecnologia, hi ha un ventall molt ampli d'estratègies i algoritmes que s'utilitzen per a crear veu sintètica a partir d'un text escrit. L'elecció de la tècnica que s'ha d'utilitzar depèn de molts factors: de la qualitat de la veu que es requereix, del nombre de veus diferents que es requereixen, de la quantitat de memòria que pot utilitzar l'aplicació, del temps de reacció del sistema…

En aquest mòdul ens centrarem en l'estructura dels sistemes de conversió de text a veu més utilitzada en els sistemes comercials actuals i en les tècniques de síntesi per concatenació en el domini temporal.

Per què s'ha d'estudiar la síntesi de la parla?

La interacció per mitjà de la parla amb els dispositius que ens envolten requereix no solament que aquests dispositius ens entenguin (reconeixement de la parla), sinó que ens puguin transmetre informació. El mitjà de transmissió d'aquesta informació pot ser divers; per exemple, visual (punts lluminosos, text, gràfiques, imatges o vídeos), acústic (sons o parla), hàptic (vibracions, temperatura, etc.) o una combinació de dos mitjans o més.

La síntesi de la parla recull totes les tècniques necessàries per a transformar en veu qualsevol informació.

Objectius

En acabar de treballar aquest mòdul, heu de ser capaços del següent:

1. Explicar el funcionament d'un convertidor de text a veu a alt nivell.

2. Identificar els mòduls principals d'un convertidor de text a veu.

3. Relacionar els conceptes de coarticulació i de discontinuïtats en els sistemes de síntesi per concatenació.

4. Calcular els nous instants i finestres de concatenació de l'algoritme TD-PSOLA per a fer modificacions prosòdiques.

1. Introducció a la síntesi de la parla

En un sentit molt ampli, la síntesi de la parla es pot definir com tot procés de **creació** de parla de manera artificial (per mitjans mecànics, electrònics, etc.).

Nota

La reproducció directa de gravacions d'àudio no es considera síntesi de la parla, ja que no hi ha associat cap procés de creació de contingut en l'instant de la reproducció.

En aquest mòdul ens centrarem en els sistemes digitals de síntesi actuals.

Bibliografia complementària

Per a una visió de l'evolució històrica dels sistemes de síntesi de la parla consulteu les referències següents:

D. H. Klatt (1987). "Review of Text-to-Speech Conversion for English". *Journal Acoustical Society of America* (núm. 82, vol. 3, pàg. 737-793).

P. Calliope (1989). *La parole et son traitement automatique* (pàg. 410-414). París: Masson / CNET-ENST ("Technique et Scientifique des Télécommunications").

R. Lingaard (1985). *Electronic Synthesis of Speech* (pàg. 1-17). Cambridge: Cambridge University Press.

Els sistemes de síntesi actuals es divideixen en dos grans grups:

- **Sistemes de resposta vocal o de text restringit**
  - Finalitat: Generació acústica de frases amb una estructura predeterminada, en què només varia part del vocabulari
  - Exemple: Avisos dels trajectes dels trens ("Pròxima estació: Palau Reial") o informació telefònica ("El número sol·licitat és 902 141 141")
  - Tècnica: Funcionament basat en la concatenació de trossos de frase o paraules que s'han gravat prèviament de manera aïllada
  - Avantatges: La qualitat de la veu generada és elevada, tot i que de vegades es noten els punts d'unió de les diferents gravacions
  - Inconvenients: No serveixen per a aplicacions en què el vocabulari és molt variable (cartelleres de cinema, per exemple), ja que incloure paraules noves en el vocabulari implica fer noves gravacions
- **Convertidors de text a veu**
  - Finalitat. Lectura de qualsevol text, amb independència de l'origen que tingui
  - Exemple. Lectors de pantalla per a persones amb deficiències visuals, entre altres

– Tècnica. Funcionament basat en dos passos: una anàlisi del text donat i la generació de l'àudio corresponent

– Avantatges. Flexibilitat, ja que tenen vocabulari il·limitat, i altament configurables (per exemple, sexe i edat del locutor, i velocitat a l'hora de parlar)

– Inconvenients. En comparació dels sistemes de resposta vocal, la complexitat tècnica és molt superior

La resta del mòdul la focalitzarem a descriure els convertidors de text a veu i les tècniques de processament d'àudio que hi estan associades.

2. Aplicacions dels convertidors de text a veu

Una part important de les aplicacions dels convertidors de text a veu o *text-to-speech conversion system* (TTS) està relacionada amb sistemes de diàleg, en què treballen juntament amb un ASR. Dins d'aquest tipus d'aplicacions destaquem l'**atenció telefònica** (serveis d'informació, de notificació d'avaries, de concertar cita, etc.) i les aplicacions de **comandament i control** (sistemes domòtics, interacció vocal amb cotxes, PC, etc.), que ja vam introduir en el mòdul "Reconeixement automàtic de la parla".

Els TTS també s'utilitzen molt en aplicacions d'**ajuda a persones amb necessitats especials**. Els lectors de pantalla, juntament amb teclats adaptats, han permès l'accés al món de la informàtica i Internet a les persones amb dificultats visuals. Per a les persones amb dificultats en la producció de la parla que els afecta la intel·ligibilitat, hi ha dispositius amb teclats predictius o de construcció ràpida de frases, que els permeten sintetitzar tot allò que volen expressar.

Un altre camp d'aplicació dels TTS és l'**aprenentatge d'idiomes**. En aquest cas, els estudiants d'idiomes estrangers poden sintetitzar qualsevol frase o paraula per aprendre una pronunciació i entonació correctes. Els sistemes de **traducció automàtica veu-veu**, tot i estar encara en fase precomercial, són un nou camp d'aplicació.

3. Els convertidors de text a veu

Els convertidors de text a veu tenen com a entrada un text i com a sortida l'àudio corresponent.

Encara que hi ha diferents tècniques per a fer aquesta conversió, tots els TTS comparteixen la mateixa arquitectura que es mostra en la figura 1:

Arquitectura TTS

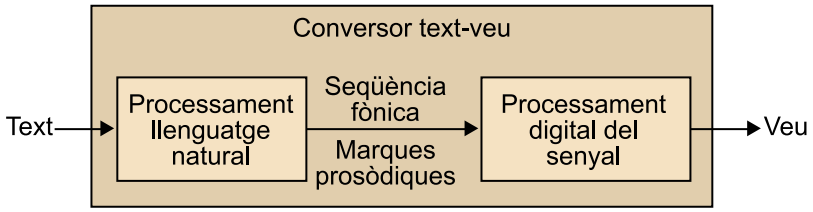


Figura 1. Arquitectura dels convertidors de text a veu

Com es veu en aquesta figura, l'arquitectura d'un TTS està formada per dos mòduls: un de processament de llenguatge natural i un altre de processament digital del senyal.

El **mòdul de processament de llenguatge natural** és l'encarregat de transformar el text d'entrada (text pla, HTML, etc.) en una transcripció fonètica amb marques prosòdiques.

Marques prosòdiques

Informació sobre quan es produeixen els canvis audibles en el to, la forma dels contorns melòdics, el volum i el ritme de la parla. Per exemple, no té la mateixa prosòdia una oració enunciativa que una d'interrogativa.

Trobareu una definició més completa de *prosòdia* en el mòdul "Introducció a la parla".

El **mòdul de processament digital del senyal** genera l'àudio descrit a l'entrada; per això també s'anomena *mòdul de síntesi acústica*.

En els propers apartats veurem amb més detall cadascun dels dos blocs d'un TTS.

3.1. El mòdul de processament de llenguatge natural

El mòdul de processament de llenguatge natural o *natural language processing* (PLN) és l'encarregat de l'anàlisi del text d'entrada i de tractar-lo perquè el pugui llegir un sistema de síntesi.

Un sistema complet de processament de llenguatge natural per a un TTS consta, majoritàriament, de tres etapes:

- Un analitzador morfosintàctic
- Un transcriptor fonètic

- Un generador prosòdic

3.1.1. Analitzador morfosintàctic

L'analitzador morfosintàctic processa el text per deixar-lo net (eliminació d'etiquetes, vinyetes, etc.) i extreu l'estructura del text, és a dir, l'organització de les frases segons les relacions de dependència que s'estableixen entre les paraules. També aporta informació sobre la categoria gramatical de les paraules o *part of speech* (POS) i normalitza el text expandint abreviacions, acrònims o transcrivint els números, entre altres operacions.

Categoria gramatical

Per *categoria gramatical* s'entén la funció que fa la paraula dins la frase. Per exemple, nom/substantiu, pronom, adjectiu, verb o adverbi.

La figura següent mostra un exemple d'anàlisi morfosintàctica per a la frase *La CE defensa el criteri de l'ONU*:

Anàlisi morfosintàctica

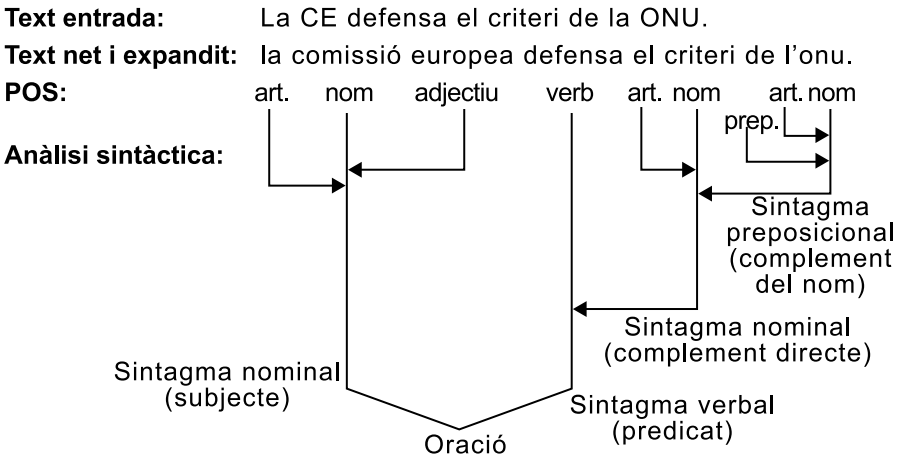


Figura 2. Exemple d'anàlisi morfosintàctica de la frase *La CE defensa el criteri de l'ONU*

Hi ha tres raons principals per les quals els TTS incorporen un analitzador morfosintàctic:

1. Per a saber com s'ha d'expandir el text no normalitzat (*CE* i *ONU* en l'exemple anterior; alguns altres casos són els números o les sigles com *PSOE*, *PP*, *IU* i *CCOO*).

2. Perquè la transcripció fonètica pot dependre de la categoria gramatical de la paraula.

3. Perquè la prosòdia depèn de la sintaxi.

3.1.2. Transcriptor fonètic

El transcriptor fonètic transforma el text net i expandit en una seqüència de fonemes, tenint en compte la informació morfosintàctica associada a cada paraula i les relacions fonètiques que s'estableixen entre unes i altres per a l'idioma escollit. S'ha de pensar, per exemple, que fer una conversió text-fo-

nemes en castellà és bastant trivial en comparació de l'anglès, ja que en castellà la relació és pràcticament 1:1 (és a dir, una lletra té en la majoria de casos només un fonema associat). Per exemple, la vocal *a* en castellà té només un fonema associat ([a]), mentre que en anglès en pot tenir diversos ([a], [ɛ], etc.).

Vegeu la transcripció en l'alfabet fonètic internacional de les frases següents:

Català: *és una bona infermera* [e'Zunəbo'niŋfərme'rə]

Castellà: *es una buena enfermera* [esu'naβwe'naeMferme'ra]

Anglès: *she is a good nurse* [ʃi' i'z ə gʊ'd nə'rs]

3.1.3. Generador prosòdic

El generador prosòdic és l'encarregat de generar de manera automàtica una descripció dels contorns melòdics i els patrons de ritme de les locucions. La prosòdia té una funció molt clara en la comunicació verbal, que és situar el focus d'atenció, transformar frases declaratives en interrogatives o exclamatives i transmetre l'estat emocional del locutor.

Informació prosòdica

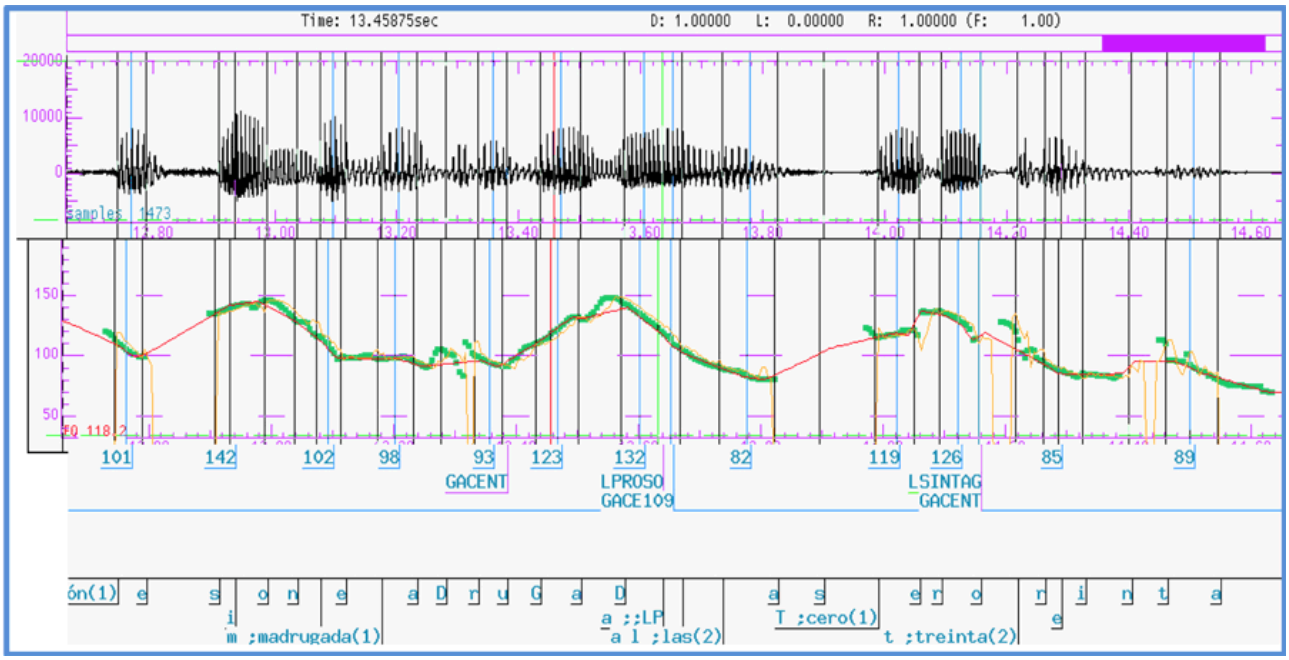


Figura 3. Exemple d'un segment de veu (gràfic superior) amb la corba melòdica (gràfic del mig) i duració de cada fonema (gràfic inferior)

Bibliografia complementària

Els mètodes utilitzats en el mòdul de processament de llenguatge natural són tècniques de lingüística computacional i intel·ligència artificial. Si en voleu més informació consulteu les referències següents:

M. A. Martí; J. Llisterri (2002). *Tratamiento del lenguaje natural: tecnología de la lengua oral y escrita*. Barcelona: Edicions de la Universitat de Barcelona.

T. Dutoit (1997). *An Introduction to Text-to-Speech Synthesis* (pàg. 37-176). Kluwer Academic Publishers.

3.2. El mòdul de processament del senyal

El mòdul de processament del senyal és l'encarregat de transformar la informació simbòlica de fonemes i marques prosòdiques, proporcionada pel mòdul de processament de llenguatge natural, en parla. Per tant, és el mòdul que genera el senyal acústic.

Les primeres estratègies de síntesi acústica que van sorgir imitaven el sistema de producció humà (síntesi articulatòria) o les característiques acústiques del senyal (síntesi per formants). En l'actualitat, la majoria de TTS comercials utilitzen la síntesi per concatenació en el mòdul de processament del senyal, per la simplicitat que té i per la qualitat de la veu que genera.

En l'apartat següent explicarem amb més detall la síntesi per concatenació.

Els TTS transformen qualsevol text en veu.

Els TTS estan formats per dos mòduls: el mòdul de processament de llenguatge natural i el mòdul de processament digital del senyal.

El mòdul de processament de llenguatge natural processa el text d'entrada per extreure'n la transcripció fonètica i genera les característiques prosòdiques més adequades per a aquest text.

El mòdul de processament digital del senyal converteix la seqüència de fonemes i la informació prosòdica, proporcionades pel mòdul de llenguatge natural, en una ona acústica.

La tècnica més utilitzada de generació de l'ona acústica és la síntesi per concatenació, tot i que hi ha altres tècniques, com per exemple la síntesi per formants i la síntesi articulatòria.

Bibliografia complementària

Per a ampliar la informació sobre els sistemes de síntesi articulatòria i síntesi per formants consulteu la referència següent:

X. Huang; A. Acero; A. Hon (2001). *Spoken Language Processing* (pàg. 796-803). Englewood Cliffs, NJ: Prentice Hall PTR.

4. Síntesi per concatenació

La síntesi per concatenació consisteix a posar l'un darrere l'altre trossos curts de gravacions d'un mateix locutor per reproduir la transcripció fonètica amb les característiques prosòdiques requerides.

Reflexió

Si les locucions se sintetitzen a partir de veu real, la qualitat dels sistemes per concatenació, és similar a la veu natural? Escoltem l'àudio d'exemple: *tts.mp3*. Clarament, la qualitat de la veu sintètica no és igual a la veu real.

La qualitat de la veu sintetitzada no és igual a la de la veu real. El principal motiu de la pèrdua de qualitat dels sistemes de concatenació és la coarticulació. El fenomen de coarticulació provoca que, quan es concatenen dos segments de veu que no eren adjacents en la gravació original, hi hagi **discontinuïtats** en els punts de concatenació, i per tant hi ha una pèrdua de qualitat/naturalitat.

Coarticulació

Adaptació del començament i final d'un so als sons adjacents, perquè la veu soni fluida. Per a més informació consulteu el mòdul "Introducció a la parla".

Hi ha dos tipus de discontinuïtats:

- **Les discontinuïtats espectrals**. Són salts bruscos en la distribució freqüencial del senyal, deguts al fet que els formants de les dues unitats no coincideixen en el punt de concatenació.

- **Les discontinuïtats prosòdiques**. Es generen principalment quan el to (f_0) o volum de cada unitat no coincideix en el punt de concatenació.

Els sistemes de síntesi per concatenació intenten minimitzar aquests dos tipus de discontinuïtats.

4.1. Elements dels sistemes de síntesi per concatenació

Els sistemes de síntesi per concatenació estan formats per una base de dades (anomenada *corpus d'unitats acústiques*) i tres blocs funcionals: un bloc de selecció d'unitats acústiques, un bloc de modificació prosòdica i un bloc de concatenació.

Sistemes de síntesi per concatenació

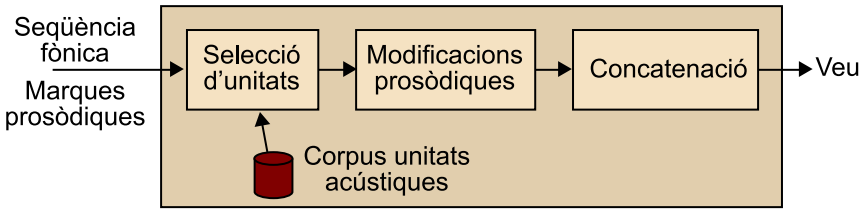


Figura 4. Mòdul de processament digital del senyal (o mòdul de síntesi acústica) d'un TTS basat en síntesi per concatenació

La generació de la veu comença quan el bloc de selecció escull els segments de veu que es concatenaran, entre tots els disponibles en el corpus d'unitats acústiques. Aquesta selecció es fa tenint en compte la transcripció fonètica i

el patró melòdic requerits. Tot seguit, es modifica la prosòdia de cada segment si és molt diferent de la prosòdia que es vol. En l'últim pas, es concatenen els segments.

El bloc de modificacions prosòdiques no és sempre en tots els TTS basats en concatenació. Si la dimensió del corpus d'unitats acústiques és prou gran, els segments seleccionats tenen una prosòdia molt semblant a la requerida, i per tant no és necessari modificar-la.

Vegeu també

Les tècniques de modificació prosòdica les introduirem en l'apartat "Modificacions prosòdiques".

4.2. Unitats acústiques i construcció del corpus

Un dels requeriments principals en l'elecció de les unitats acústiques que s'han d'utilitzar en un TTS és que es generin poques discontinuïtats en la concatenació.

Una de les unitats més utilitzades en els TTS són els **difonemes**, ja que tenen el començament i final en zones estables acústicament i, per tant, proporcionen una distorsió en la concatenació més baixa que altres unitats.

Difonema

Recordeu que ja hem vist la definició de difonema en el mòdul "Introducció a la parla".

Un difonema és la unitat acústica que comença al mig de la zona estable d'un fonema i acaba al mig de la zona estable del fonema següent.

Per a construir el corpus s'ha de dissenyar un conjunt de frases que incloguin totes les unitats acústiques que es voldran sintetitzar i fer gravacions amb un locutor professional. Un cop enregistrat el material, s'etiqueten totes les gravacions juntament amb la informació prosòdica que hi està associada (l'altura tonal, la duració, l'energia, etc.). Finalment, si la unitat acústica que utilitzarem és el difonema, s'han d'indicar els límits de cada segment de senyal pertanyent a un difonema per a guardar-lo al corpus juntament amb la informació prosòdica (altura tonal, duració, etc.).

El procés de creació del corpus d'un TTS requereix molta intervenció manual: l'elecció del locutor, els enregistraments i la revisió dels etiquetatges. El procés d'etiquetatge és una tasca semiautomàtica; això significa que hi ha una primera fase d'etiquetatge automàtic i una segona fase de revisió manual.

Veu d'un TTS

El locutor que enregistra el corpus d'unitats lingüístiques *cedeix* la veu al TTS. S'ha de dir que no tothom pot ser locutor d'un corpus, ja que no tots els locutors generen veus sintètiques de bona qualitat. Per aquest motiu, es contracten locutors professionals.

4.3. Selecció de les unitats

L'objectiu d'aquest bloc és escollir la successió d'unitats acústiques òptima per a sintetitzar la descripció fonètica proporcionada i que, a més a més, s'ajusta a la prosòdia volguda.

L'elecció de les unitats per a una seqüència de fonemes no és trivial, ja que en la base de dades normalment hi ha més d'una instància per a cada unitat i cadascuna de les quals té una prosòdia diferent. La tècnica més utilitzada per a fer aquesta selecció és una cerca utilitzant l'algoritme de Viterbi que minimitza una funció de cost que representa la distorsió introduïda.

La **funció de cost** està formada per dos termes:

- El **cost d'unitat** representa la diferència entre els valors de prosòdia i de context que té la unitat escollida respecte als valors volguts. Normalment es calcula proporcionalment a la diferència entre les característiques prosòdiques (f_0 o $\log(f0)$, duració, energia, etc.) de la unitat escollida i la descripció volguda d'entrada al mòdul.

- El **cost de concatenació** representa la pèrdua de qualitat deguda a la unió de dues unitats. Es calcula com la distància entre els paràmetres espectrals de dues unitats adjacents.

La cerca prioritza la selecció d'unitats adjacents en les gravacions originals, sempre que les diferències prosòdiques no siguin gaire notables. Per tant, es redueixen els punts de concatenació i s'eviten pèrdues de qualitat.

4.4. Concatenació

La concatenació és l'últim pas per a sintetitzar una locució. L'objectiu que té és unir la seqüència de segments acústics seleccionats prèviament de manera que disminueixin les discontinuïtats que hi ha en els punts de concatenació.

Per a reduir les discontinuïtats s'utilitzen tècniques de suavització. Segons el domini en què s'apliquen, es distingeix entre concatenació en el domini temporal i concatenació en el domini paramètric.

4.4.1. Concatenació en el domini temporal

La concatenació en el domini temporal es basa en la idea de superposar i sumar o, en anglès, *overlap-and-add* (OLA). Aquesta tècnica s'aplica en dues fases: l'anàlisi (quan es defineixen els segments de veu que se superposaran) i la síntesi (quan es fa la suma).

En la fase d'anàlisi es construeixen finestres de veu multiplicant els segments acústics seleccionats per finestres superposades. Els instants en què es fan aquestes multiplicacions s'anomenen *instants d'anàlisi* t_a (vegeu la part superior de la figura 5).

Una pràctica habitual, anomenada *time domain pitch synchronous overlap-and-add* (TD-PSOLA), és situar els punts d'anàlisi separats un període d'altura tonal i escollir finestres d'anàlisi de durada dos períodes d'altura tonal.

En la fase de síntesi, els segments de veu se situen en els instants de síntesi t_s que els corresponen (depenent de quina altura tonal té la veu que volem crear) i se sumen (vegeu la part inferior de la figura 5).

La figura 5 mostra un exemple TD-PSOLA per a la concatenació de dues unitats acústiques:

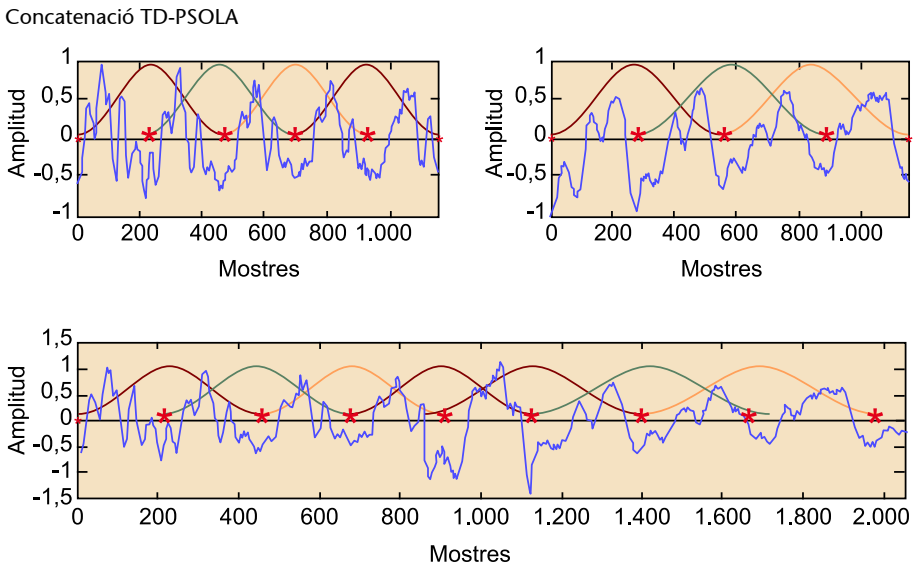


Figura 5. Concatenació TD-PSOLA de dues unitats acústiques. A dalt: els dos segments temporals que s'han de concatenar. A baix: el resultat de la concatenació. Els punts d'anàlisi i síntesi s'han indicat amb un asterisc vermell.

En els intervals al voltant dels punts de concatenació, el valor de la forma d'ona resultant és la interpolació lineal de les dues unitats acústiques.

4.4.2. Concatenació en el domini paramètric

Per a fer la concatenació en el domini paramètric primer s'han de transformar els segments acústics seleccionats prèviament a la representació paramètrica segons el model de veu utilitzat. Els models de veu més utilitzats per a la concatenació són dos: el model font-filtre, i el model harmònic o l'extensió harmònic + soroll d'aquest model.

Perquè la transició entre els dos segments acústics que s'han de concatenar sigui suau, a la frontera entre els dos segments s'interpolen els paràmetres del model de veu de cada segment. Finalment, a partir dels nous paràmetres interpolats es genera la veu.

La síntesi per concatenació consisteix a crear la veu sintètica mitjançant la unió de segments curts pregravats d'un mateix locutor.

Tots els esforços dels sistemes de concatenació estan centrats a minimitzar les discontinuïtats espectrals i prosòdiques en els punts de concatenació.

Els principals components dels sistemes de concatenació són els següents:

- Un corpus d'unitats acústiques. Les unitats més utilitzades són les que tenen el començament i el final en zones estables freqüencialment (per exemple, els difonemes).

- Un bloc de selecció de les unitats que minimitza el cost d'unitat (diferència entre la unitat seleccionada i l'especificada) i el cost de concatenació (mesura de les discontinuïtats introduïdes per la unitat en els punts de concatenació).

- Un modificador de la prosòdia (bloc opcional, depenent de la dimensió del corpus).

- Un bloc de concatenació en el domini temporal o paramètric.

Les tècniques de concatenació més habituals en el domini temporal es basen en l'algoritme OLA.

Bibliografia complementària

Si voleu profunditzar en els models de veu i la utilització d'aquests models en la síntesi per concatenació consulteu les referències següents:

T. Dutoit (1997). *An Introduction to Text-to-Speech Synthesis* (pàg. 229-248). Kluber Academic Publishers.

X. Huang; A. Acero; A. Hon (2001). *Spoken Language Processing* (pàg. 290-333). Englewood Cliffs, NJ: Prentice Hall PTR.

5. Modificacions prosòdiques

Les tècniques de modificació prosòdica permeten modificar la duració, l'altura tonal i l'energia de segments de veu, sense modificar el missatge. Aquestes tècniques es fan servir en el bloc de modificació prosòdica d'un TTS, si en té (recordem que és opcional). També es poden utilitzar per a canviar la prosòdia de qualsevol gravació de veu original; per exemple, per a ajustar la duració a un valor requerit o per a ajustar les característiques de la veu a uns valors volguts (per exemple, per a apujar o abaixar el to de veu o el volum).

Per a modificar l'energia de la veu, és a dir, el volum de la veu, només s'ha de multiplicar el senyal per un valor volgut (anomenat també *guany*). En canvi, la modificació de la duració i l'altura tonal requereixen manipular el senyal i la qualitat de la veu resultant es degrada.

Hi ha, igual que per a la concatenació, dues aproximacions per a modificar l'altura tonal i la duració: tècniques basades en la modificació dels paràmetres del senyal de veu i tècniques basades en OLA. Aquestes últimes tècniques són de les més utilitzades, sobretot les basades en TD-PSOLA, ja que són senzilles d'implementar i generen una veu modificada de qualitat acceptable.

En l'apartat següent descrivim el funcionament de les tècniques de modificació prosòdica basades en TD-PSOLA.

5.1. Tècniques basades en TD-PSOLA

L'algoritme de modificació prosòdica en el domini temporal TD-PSOLA o *time domain pitch synchronous overlap-and-add* consisteix a modificar els instants de síntesi i la seqüència de finestres que s'han de concatenar depenent del tipus de modificació que es vulgui.

5.1.1. Modificacions d'altura tonal

En l'algoritme TD-PSOLA, les finestres d'anàlisi i síntesi se situen separades 1 període d'altura tonal. Per tant, si es vol modificar l'altura tonal només s'han de resituar els instants de síntesi segons el nou període d'altura tonal. En la figura 6 es mostra un exemple de modificació d'altura tonal per a un factor d'1,25 (multiplicació de 1 / 1,25 = 0,8 del període d'altura tonal).

Modificació d'altura tonal basada en TD-PSOLA

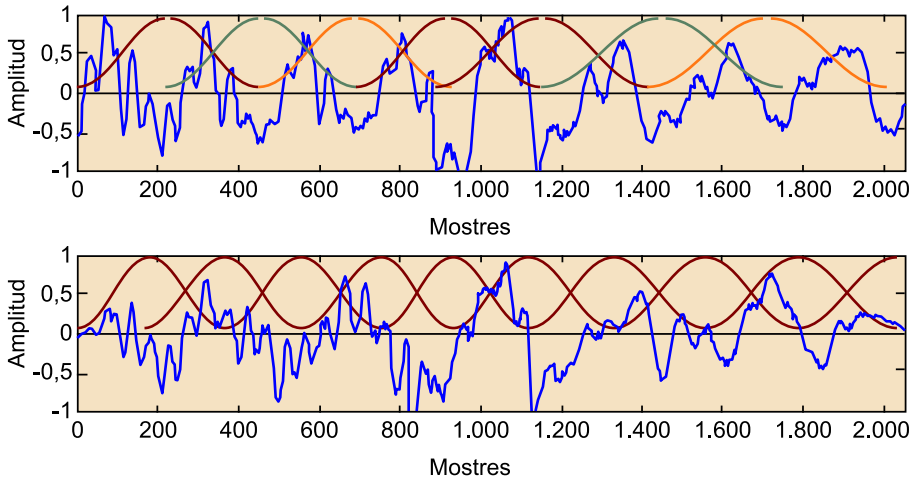


Figura 6. Modificació de TD-PSOLA d'altura tonal per un factor constant d'1,25

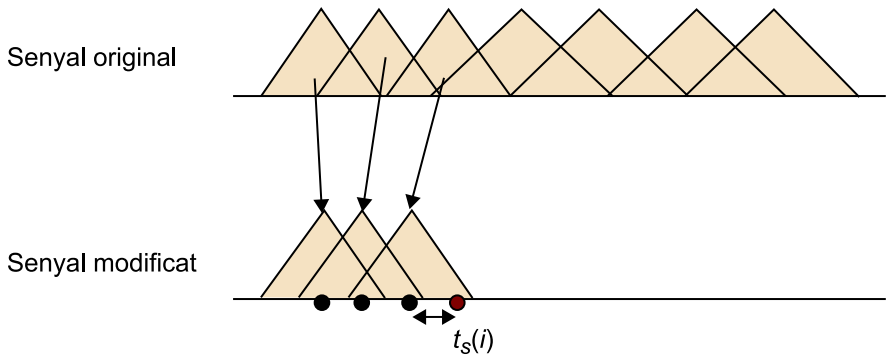
Escolteu un exemple per a valorar la qualitat i l'efecte de la modificació d'altura tonal: "locució original" i "locució modificada".

Com observem en la figura anterior, perquè no hi hagi modificació de la duració total del segment de veu en la síntesi es repeteixen o s'eliminen finestres d'anàlisi.

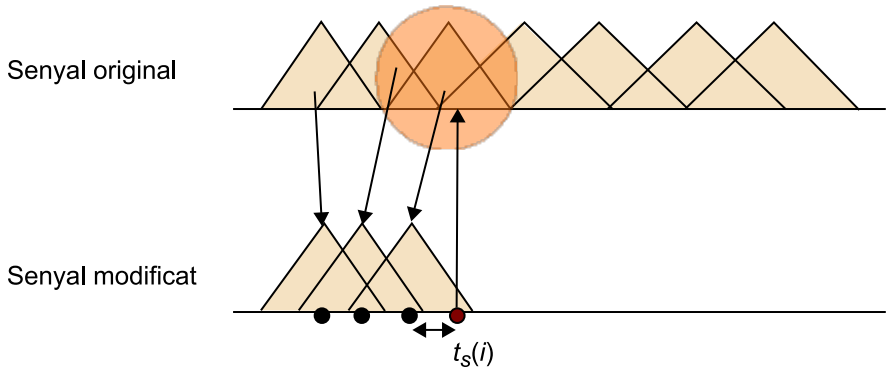
D'una manera més esquemàtica, l'algoritme de TD-PSOLA per a modificar altura tonal consisteix en els passos que presentem a continuació. Hem il·lustrat cada pas de l'algoritme amb un petit exemple gràfic en què representem de manera esquemàtica un senyal original que s'ha de modificar i hi indiquem les finestres d'anàlisi i el senyal modificat que es va generant:

1. Determinar el nou instant de síntesi $t_s(i)$ segons el factor de modificació $\alpha(i)$, del període d'altura tonal $T_{pitch}(i)$ i de l'instant de síntesi previ $t_s(i-1)$:

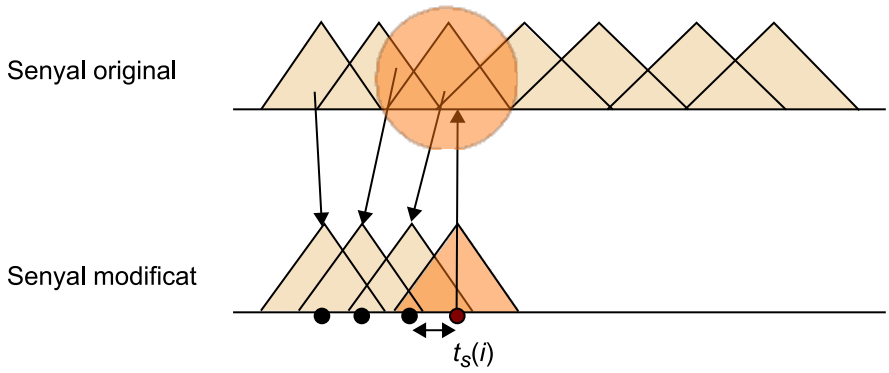
$$t_s(i) = t_s(i-1) + T_{pitch}(i) * \alpha(i)$$



2. Seleccionar la finestra d'anàlisi més propera a l'instant de síntesi:



3. Situar la finestra d'anàlisi seleccionada en l'instant de síntesi:



4. Un cop s'han situat totes les finestres en els instants de síntesi corresponents, s'ha de fer la suma. En aquest exemple hem reduït el període d'altura tonal; per tant, el so modificat és més agut que l'original.

Els canvis d'altura tonal que permet l'algoritme de TD-PSOLA estan entre valors de 0,5 (reduir a la meitat) i 2 (doblar), perquè si no la qualitat queda molt degradada i es pot perdre el missatge del segment de veu.

Per a les parts sordes del senyal de veu, no es fa cap modificació.

5.1.2. Modificacions de duració

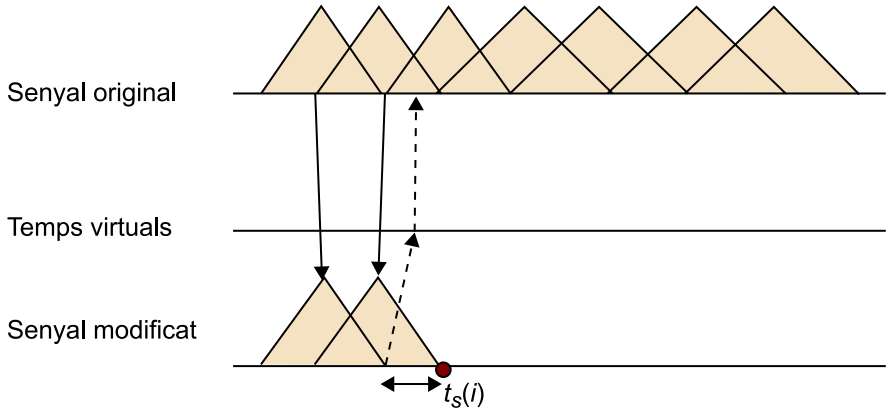
Les modificacions de duració es fan repetint o eliminant finestres d'anàlisi, mantenint sempre l'altura tonal original.

Escolteu la "locució original" i la "locució modificada" per a valorar la qualitat i l'efecte d'una modificació de duració per un factor constant d'1,25.

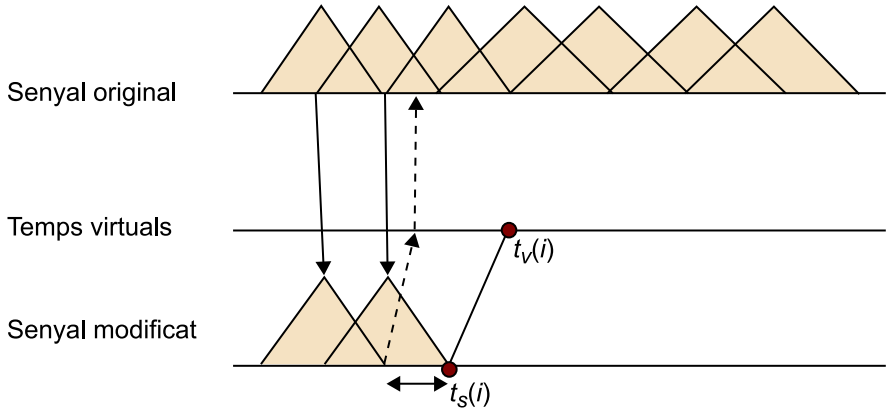
Per a seleccionar les finestres que se situen en cada punt de síntesi s'ha de construir una seqüència d'instants virtuals. Els passos de l'algoritme són els que descrivim a continuació. Com abans, hem il·lustrat cada pas de l'algoritme amb un petit exemple gràfic en què representem de manera esquemàtica un senyal original que s'ha de modificar i hi indiquem les finestres d'anàlisi i el senyal modificat que es va generant:

1. Determinar el nou instant de síntesi $t_s(i)$ segons el període d'altura tonal $T_{pitch}(i)$ i de l'instant de síntesi previ $t_s(i-1)$:

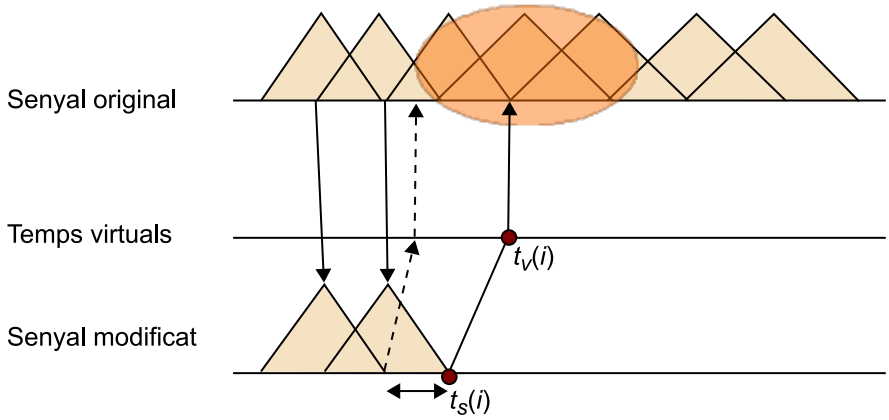
$$t_s(i) = t_s(i-1) + T_{pitch}(i)$$



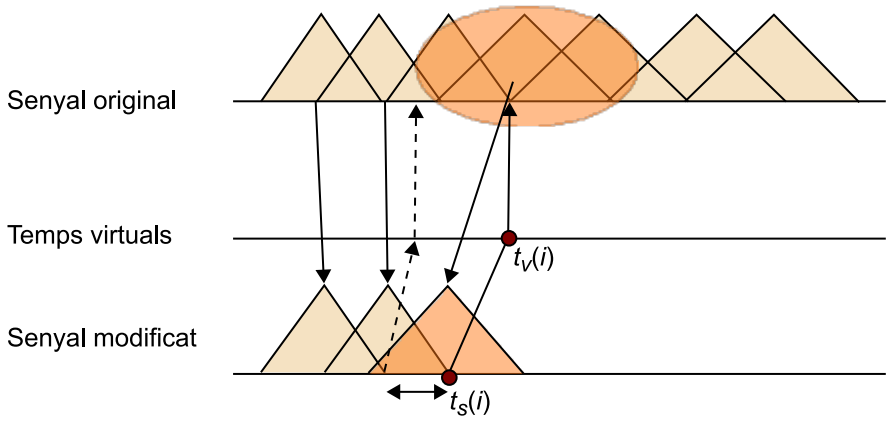
2. Determinar el nou instant virtual $t_v(i)$ segons l'instant de síntesi $t_s(i)$ i el factor de modificació de duració $\beta(i)$: $t_v(i) = t_s(i-1) + (t_s(i) - t_s(i-1)) * \beta(i)$:



3. Seleccionar la finestra d'anàlisi més propera a l'instant virtual:



4. Situar la finestra d'anàlisi seleccionada en l'instant de síntesi:



5. Un cop s'han situat totes les finestres en els instants de síntesi corresponents, s'ha de fer la suma. En aquest exemple hem reduït la durada del senyal original.

Els canvis de duració que permet l'algoritme de TD-PSOLA estan entre valors de 0,25 i 2, perquè si no la qualitat queda molt degradada i es pot perdre informació rellevant per al missatge transportat per la veu.

Per a les parts sordes del senyal de veu, els instants d'anàlisi i síntesi se situen en intervals regulars (habitualment d'uns 10 mil·lisegons).

5.1.3. Alternatives a l'algoritme de TD-PSOLA

L'estimació dels instants d'anàlisi síncrons amb l'altura tonal és un procediment complicat i poc precís, que provoca desajustos de fase, altura tonal i amplitud en els punts de concatenació de la síntesi i, per tant, efectes audibles en el senyal modificat.

Per aquesta raó van sorgir tot un conjunt de tècniques de la família PSOLA que intenten minimitzar les discontinuïtats en els punts de concatenació, com per exemple FD-PSOLA (*frequency domain pitch synchronous overlap-and-add*) i MBR-PSOLA (*multiband resynthesis pitch synchronous overlap-and-add*).

Bibliografia complementària

Si voleu profunditzar en altres algoritmes de modificació prosòdica basats en PSOLA consulteu les referències següents:

T. Dutoit (1997). *An Introduction to Text-to-Speech Synthesis*. Kluber Academic Publishers.

Joc d'eines de codi obert per a construir un TTS:

Festival

6. Mesures de qualitat de la veu sintetitzada

Per a saber quina és la qualitat de la veu generada per a un TTS, la resposta que sembla més evident és escoltar la veu sintetitzada.

Exemple

Per a tenir una idea de la qualitat actual dels TTS consulteu els exemples comercials següents:

Loquendo

Verbio

Nuance

Tanmateix, escoltant-ne només un exemple no es pot avaluar la qualitat d'un TTS, ni fer una comparació entre diferents sistemes, ja que els TTS no es comporten igual amb totes les frases d'entrada.

Els tests més utilitzats per a avaluar els TTS són els següents:

- Test d'intel·ligibilitat. Els avaluadors escolten monosíl·labs i han de marcar quin han entès d'una llista d'opcions molt semblants acústicament (per exemple: *mou*, *nou*). També es demana als avaluadors que facin transcripcions de frases sintetitzades sense sentit (per exemple, *la pilota menja tres cotxe amunt*).

- Test de qualitat o *mean opinion score* (MOS). S'avalua la qualitat de cada locució de l'1 al 5 (1 = *bad*; 2 = *poor*; 3 = *fair*; 4 = *good*; 5 = *excellent*).

Per a comparar diferents TTS entre si s'ha d'assegurar que els mateixos avaluadors han passat els mateixos tests.

Els TTS comercials actuals tenen una qualitat prou bona per a ser acceptats pel mercat per a la majoria d'aplicacions que requereixen una veu neutra. Avui en dia, els principals esforços de recerca i comercialització estan centrats en dos aspectes: la reducció de requeriments de memòria per a poder introduir un TTS en qualsevol dispositiu i la introducció d'expressivitats (per exemple, emocions o riure) en la veu sintetitzada.

Els TTS transformen qualsevol text en veu.

Els TTS estan formats per dos mòduls: el mòdul de processament de llenguatge natural (encarregat de generar la transcripció fonètica i les marques prosòdiques) i el mòdul de processament digital del senyal (encarregat de generar l'ona acústica).

La tècnica més utilitzada de generació de l'ona acústica és la síntesi per concatenació, que consisteix a unir segments curts pregravats d'un mateix locutor per a formar locucions. Tots els esforços dels sistemes de concatenació estan centrats a minimitzar les discontinuïtats espectrals i prosòdiques en els punts de concatenació.

Un dels algoritmes més utilitzats per a fer la concatenació i les modificacions prosòdiques de la veu és el TD-PSOLA. En la fase d'anàlisi, se segmenta la veu original en finestres encavalcades de longitud dos cops el període d'altura tonal i alhora situades a cada període d'altura tonal. En la fase de síntesi es construeix el senyal sumant les finestres encavalcades que hi ha a cada instant de síntesi.

La qualitat de la veu generada per un TTS és prou bona perquè hi hagi múltiples aplicacions comercials.