ISSN (electrònic): 2014-2242 / www.cac.cat

QUADERNS DEL CAC

Aplicació de tècniques de generació automàtica de la parla en producció audiovisual

Francesc Alías
Membre del Grup de Recerca en Tecnologies Mèdia de La Salle - Universitat Ramon Llull

falias@salle.url.edu

Ignasi Iriondo
Membre del Grup de Recerca en Tecnologies Mèdia de La Salle - Universitat Ramon Llull

iriondo@salle.url.edu

Joan Claudi Socoró
Membre del Grup de Recerca en Tecnologies Mèdia de La Salle - Universitat Ramon Llull

jclaudi@salle.url.edu

Article rebut el 01/06/2011 i acceptat el 13/12/2011

Resum

En aquest article es presenta un resum del treball de recerca que porta el mateix títol, realitzat gràcies a l'ajut concedit pel CAC en la VII convocatòria d'Ajuts a projectes de recerca sobre comunicació audiovisual. Després d'estudiar el grau d'implantació dels sistemes de síntesi de veu a Catalunya, se n'analitza la viabilitat de l'ús en l'àmbit de la creació de produccions audiovisuals. En aquest article es presenten les conclusions de l'estudi de camp realitzat i dels experiments desenvolupats a partir del sistema de síntesi de la parla de La Salle (Universitat Ramon Llull) adaptat al català.

Paraules clau

síntesi de veu, produccions audiovisuals, audiodescripció, valoració subjectiva de qualitat.

Abstract

This article presents a summary of the research work of the same title, developed thanks to the grant awarded by the CAC in the VII call of research projects on audiovisual communication. After studying the degree of implementation of speech synthesis systems in Catalonia, we analyze the feasibility of its use for the creation of audiovisual productions. This article presents the findings of the field study and the experiments developed after adapting the speech synthesis system of La Salle (Universitat Ramon Llull) to the Catalan language.

Keywords

speech synthesis, audiovisual productions, audio description, subjective assessment of quality.

1. Introducció

La síntesi de veu o de la parla és la tècnica que permet generar automàticament una locució amb característiques similars a les d'una veu humana a partir d'un text d'entrada. Els sistemes de síntesi de veu es poden arribar a confondre amb els sistemes que fan un ús de veu gravada per a la reproducció de missatges de veu, però cal tenir clar que, en general, la síntesi de veu es refereix a les tècniques que permeten generar qualsevol missatge oral.

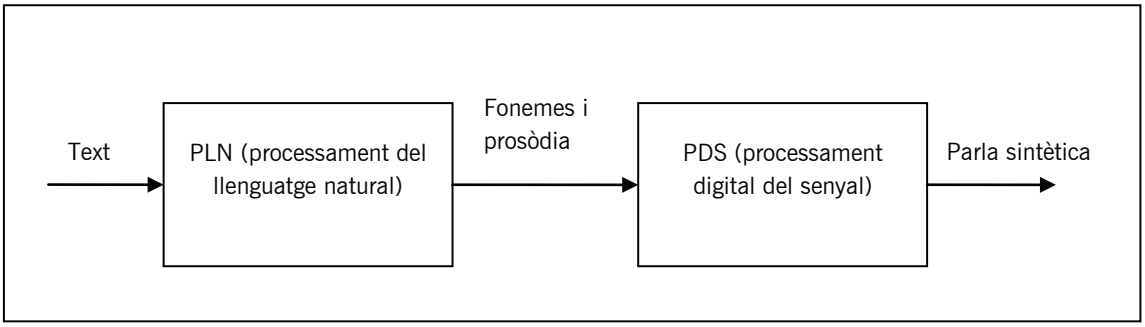
El text d'entrada pot provenir d'un correu electrònic, d'un web o bé es pot escriure directament des d'un teclat. Algunes de les aplicacions típiques d'aquest tipus de sistemes són l'ajuda a persones amb una determinada discapacitat (per exemple, visual), el suport per a l'aprenentatge de llengües, les aplicacions telefòniques, les aplicacions multimèdia i les interfícies persona-màquina en general.

Lluny de voler imitar el procés real amb què els humans generem la parla, existeix un model funcional que permet abordar, amb els recursos disponibles avui dia, la construcció d'un sistema que converteixi un text d'entrada qualsevol en la seva veu sintètica corresponent. Aquest model funcional, estès i acceptat àmpliament per la comunitat dedicada a les tecnologies de la parla, és el que es descriu en el diagrama de blocs de la figura 1.

Com es pot observar a la figura 1, en primer lloc tenim el bloc de processament del llenguatge natural de la parla (PLN), que és l'encarregat de trobar, a partir del text d'entrada que es vol "llegir", quina és la transcripció fonètica del text (és a dir, quins son els sons que s'han de produir al llarg de la locució de sortida) i també quina ha de ser la prosòdia associada (com han de sonar cadascun d'aquests sons, específicament sobre les seves característiques tant d'entonació com de ritme). En segon lloc, apareix el bloc de processament digital del senyal (PDS), que s'encarrega de generar, a partir dels requeriments donats pel mòdul anterior, el senyal de parla sintètica de sortida.

La resta d'aquest article està estructurat de la forma següent: la secció 2 presenta un recull d'entitats d'arreu que destaquen per la seva aportació en el món de la síntesi de veu. La secció 3 mostra els resultats més representatius d'un estudi de camp sobre la síntesi de veu en l'entorn de l'audiovisual a Catalunya,

Figura 1. Model funcional d'un sistema de conversió de text en parla (CTP)



Font: Elaboració pròpia.

com també en col·lectius de persones amb capacitats visuals reduïdes. Mitjançant un conjunt d'entrevistes personals, s'han recollit les opinions més rellevants en relació amb el grau de maduresa assolit per aquesta tecnologia, les limitacions que més destaquen del seu ús i els reptes de futur que han de permetre en un futur un grau més elevat de penetració de la conversió de text a parla en els sectors esmentats. La secció 4 descriu el procés d'adaptació del sintetitzador de veu de La Salle (URL) al català, seguint un dels objectius fixats en el projecte de recerca finançat. Utilitzant aquest sintetitzador, s'han realitzat les proves que descriu la secció 5, les quals han permès validar de forma subjectiva la viabilitat de l'ús de la síntesi de veu com a eina per generar material audiovisual (concretament, amb exemples d'anuncis i notícies). Finalment, la secció 6 inclou les conclusions principals d'aquest treball i les línies d'investigació que poden permetre avançar cap a un grau més elevat d'implantació de la síntesi de veu en els mitjans audiovisuals.

2. Implantació de la síntesi de veu en el món audiovisual a Catalunya

Per tal d'estudiar el grau d'implantació real de les tecnologies de síntesi de veu a Catalunya en el món de l'audiovisual, s'ha realitzat un treball de camp extens per tal de recollir les opinions dels seus actors principals davant la implantació actual i la possible introducció futura dels sistemes de síntesi de veu en els mitjans de comunicació audiovisual. A més, durant aquest procés s'ha pogut constatar que hi ha una part de la població, les persones amb discapacitat visual, que són grans consumidors dels sistemes de síntesi de veu. Per aquest motiu, aquest grup d'usuaris també s'ha inclòs en l'estudi realitzat per tal de conèixer la seva opinió respecte a l'ús de les tecnologies de síntesi de la parla en el context de les produccions audiovisuals (Torrens 2010).

A continuació, es presenta un recull representatiu de les empreses, els centres de recerca i els productes més rellevants en el context de la generació de veu sintètica en català. En aquest context, es recullen tant empreses d'àmbit català com internacional, així com productes que es troben a la xarxa.

2.1 Universitats i centres de recerca

1) TALP (Tecnologies i Aplicacions del Llenguatge i la Parla) de la Universitat Politècnica de Catalunya

Quant a la síntesi de veu en català cal destacar, per una banda, que el grup de recerca TALP disposa d'un sistema propi de conversió de text a parla, anomenat OGMIOS (*http://www.talp.cat/ttsdemo/index.php*), i, per l'altra, que van treballar en la incorporació del català a la plataforma Festival, desenvolupada pel sistema operatiu Linux (*http://www.cstr.ed.ac.uk/projects/festival/*), i el resultat és FestCat, que es va incloure en la distribució Linkat de la Generalitat de Catalunya. Totes aquestes aplicacions es poden descarregar gratuïtament des del web de FestCat i es publiquen sota els termes de la llicència LGPL. Per a més informació, consulteu el lloc web <http://gps-tsc.upc.es/veu/festcat/>.

Part d'aquest treball es va desenvolupar en el marc del projecte Tecnoparla: Tecnologies de la parla en català, enfocat a estudiar la viabilitat de la traducció de veu aplicada a la traducció de notícies audiovisuals. El projecte va estudiar les diferents tecnologies clau que intervenen en un sistema de traducció de veu (reconeixement, traducció i síntesi de veu), es va centrar en la incorporació del català i va abordar el progrés en les tres tecnologies implicades i la seva integració. Concretament, pel que fa a la síntesi de veu es va utilitzar el sistema de programari obert Festival (Linux) adaptat al català (FestCat). Podeu trobar més informació al lloc web següent: <http://www.talp.cat/tecnoparla/>

2) GTM (Grup de Recerca en Tecnologies Mèdia), La Salle - Universitat Ramon Llull

Aquest grup té una àmplia experiència en el món de la generació de la parla sintètica. Des dels seus inicis, al final dels anys vuitanta, ja es va centrar en la recerca en síntesi de la parla en català, mitjançant treballs de Martí (1985) i Camps (1992), posteriorment continuats per Guaus i Iriondo (2000) i Iriondo *et al.* (2004), aquest darrer treball enfocat en la síntesi expressiva (emotiva) en català.

Per a més informació, podeu consultar el lloc web següent: <http://www.salle.url.edu/portal/departaments/home-depts-DTM-projectes-PM?cf_seccio=PM&pag=1>

3) Barcelona Media - Centre d'Innovació de la Fundació Barcelona Media

Barcelona Media incorpora una línia de recerca en veu i llenguatge que centra la investigació en el processament del llenguatge, tant escrit com oral, i desenvolupa aplicacions en correcció i traducció automàtiques, anàlisi i processament de la informació, generació automàtica de textos a partir de bases de dades, i síntesi de veu, per tal de disposar d'eines per al processament automatitzat de continguts lingüístics en entorns multilingües o en què el llenguatge humà es converteix en la modalitat d'interacció prioritària.

En l'àmbit de la síntesi de veu treballen amb l'objectiu de crear una veu sintètica catalana, una de castellana i una de bilingüe (catalana i castellana), així com d'introduir-hi naturalitat expressiva i entonativa (prosòdia) i facilitar la creació de locutors especialitzats. Podeu trobar més informació al lloc web següent: <http://www.barcelonamedia.org/linies/7/ca>.

2.2 Empreses

1) Verbio

Empresa dedicada a vendre productes relacionats amb les tecnologies de la parla ubicada a Barcelona.

- Quant a la síntesi de la parla, ofereixen conversió de text a parla en diferents idiomes.
  <http://www.verbio.com/webverbio3/html/productes.php?id=1>
- Demostracions de les veus en català: Meritxell i Oriol.
  <http://www.verbio.com/webverbio3/html/demos_ttsonline.php>
- Demostracions de notícies:
  <http://www.verbio.com/webverbio3/html/demos_news.php>
  Enllaça a Vilaweb.cat, però indica que no hi ha notícies disponibles.

2) Loquendo

Empresa dedicada a la venda de productes relacionats amb les tecnologies de la parla.

- Quant a la síntesi de la parla, ofereixen conversió de text a parla en diferents idiomes. Es tracta d'un sistema de síntesi de veu basat en selecció d'unitats.
  <http://www.loquendo.com/es/technology/tts.htm>
- Demostracions de les veus en català: Montserrat i Jordi.
  <http://www.loquendo.com/es/demos/demo_tts.htm>

3) CereProc

L'empresa CereProc, en col·laboració amb Barcelona Media, ha desenvolupat un sistema de síntesi de veu femenina bilingüe en català i en castellà. Ofereixen una veu femenina sintètica, bilingüe, en català i en castellà, amb entonació natural, disponible per a múltiples aplicacions. El projecte ha comptat amb el suport de la Generalitat de Catalunya.

<http://www.cereproc.com/products/voices>

4) Nuance

Nuance Vocalizer (abans RealSpeak) disposa d'una veu femenina en català (Núria). Tanmateix, no es pot trobar massa informació al lloc web de l'empresa.

<http://www.nuance.es/realspeak/>

<http://www.nuance.com/for-business/by-solution/contact-center-customer-care/cccc-solutions>

<services/vocalizer/vocalizer-languages/index.htm>

5) Telefónica I+D

Disposa d'un sistema de conversió de text en parla multilingüe (Rodríguez *et al.* 2008). No s'ha trobat informació que ens permeti afirmar que es tracta d'un producte independent que ofereix l'empresa (vegeu *http://www.tid.es)*. Tanmateix, és una tecnologia que l'empresa ha incorporat a algun dels seus productes, com ara el lector de missatges curts *(http://saladeprensa.telefonica.es/documentos/24moviles.pdf)* o per ajudar persones amb discapacitat *(http://saladeprensa.telefonica.es/documentos/22comunicador.pdf).*

2.3 Altres productes

1) eSpeak

eSpeak és un sistema de síntesi basat en formats que treballa sota les plataformes Linux i Windows, i que es pot emprar sota la llicència GNU *General Public License* (programari lliure).

<http://espeak.sourceforge.net/>

2) JAWS (*Job Access With Speech*)

Està dirigit a persones cegues o de baixa visió.

- Llegeix el contingut de la pantalla mitjançant veu sintètica.
  <http://www.freedomscientific.com/products/fs/jaws-product-page.asp>
- Incorpora la veu en català pel fet que incorpora sistemes de síntesi d'altres empreses, com pot ser Nuance (Núria).
  <http://www.freedomscientific.com/downloads/jaws/JAWS10-whats-new.asp>

3. Implantació de la síntesi de veu en el món audiovisual a Catalunya

Per tal d'estudiar el grau d'implantació real de les tecnologies de síntesi de veu a Catalunya en el món de l'audiovisual, s'ha realitzat un treball de camp extens per tal de recollir les opinions dels seus actors principals davant la implantació actual i la possible introducció futura dels sistemes de síntesi de veu en els mitjans de comunicació audiovisual. A més, durant aquest procés s'ha pogut constatar que hi ha una part de la població, les persones amb discapacitat visual, que són grans consumidors dels sistemes de síntesi de veu. És per això que aquest grup d'usuaris també s'ha inclòs en l'estudi realitzat per tal de conèixer la seva opinió respecte a l'ús de les tecnologies de síntesi de la parla en el context de les produccions audiovisuals. Els detalls del treball de camp es poden trobar a Torrens (2010).

A continuació, s'analitzen els resultats obtinguts del treball de camp a partir de les diferents respostes recollides de les entrevistes que s'han realitzat als actors principals del sector (emissores de ràdio, televisió, productores i estudis de so i doblatge), per mitjà d'entrevistes realitzades a persones que treballen en aquest sector, tant des del vessant tècnic com del no tècnic.

A més a més, també s'ha entrevistat un grup d'usuaris potencialment molt interessat en la inclusió de la síntesi de veu en el món de la comunicació audiovisual, com és el de les persones amb discapacitat visual. Tot seguit es presenten les conclusions de l'estudi contextualitzades per aquest sector de la societat.

3.1 Mitjans de comunicació

Les entrevistes realitzades als mitjans de comunicació audiovisual s'han desglossat en tres grans grups: 1) ràdios, 2) televisions i productores de televisió, i 3) estudis d'àudio, de doblatge i de postproducció. Es va contactar amb la gran majoria d'entitats líders del sector dins del territori català, ja que l'estudi se centrava en l'aplicació de la síntesi de veu en català. De totes les entitats esmentades, van atendre l'entrevista les 19 entitats següents mitjançant un representant dels seus departaments tècnics/d'emissions (vegeu Torrens 2010 per a més detalls):

1. Ràdios: Catalunya Ràdio, 40 Principales Barcelona, COM Ràdio, RAC 1 i Onda Rambla - Punto Radio.
2. Televisions i productores: TV3, 8tv, RAC105tv i Gestmusic.
3. Estudis d'àudio, de doblatge i de postproducció: Oido (*http://www.oido.net/*), INFINIA (*http://www.infinia.es/*), Onda Estudios (*http://www.ondaestudios.com/*), Cyo Studios (*http://www.cyostudios.com/*), Dubbing Films (*http://www.dubbingfilms.com/*), Tatudec (*http://www.tatudec.com/*), Dvmusic (*http://www.dv-music.es/*), Seimar RLM Estudios, Soundub (*http://www.abaira.es/*) i Sonygraf (*http://www.sonygraf.com*).

D'aquestes entrevistes, se'n pot concloure el següent:

- Tant les ràdios i les televisions com els estudis de so són coneixedors de la tecnologia dels sistemes de síntesi de veu.
- Analitzant el primer dels grups, cap de les emissores de ràdio amb les quals s'ha contactat utilitza els sistemes de síntesi de veu, llevat d'un parell que l'han usat només per generar veu robòtica o per crear algun efecte concret, i ho han fet utilitzant programari lliure.

Hi ha diverses opinions respecte a l'ús de les tecnologies de síntesi de la parla en un futur: dues de les persones representants dels departaments tècnics de les emissores creuen que podrien ser útils, però només de manera complementària, és a dir, per a la creació d'efectes o per a emissores automatitzades. Una altra exposa que es perdria l'encant i la màgia que dóna un mitjà com la ràdio; les dues restants pensen que els sintetitzadors encara es troben lluny de ser utilitzats per la manca d'expressió i d'entonació natural en la veu.

- En cap de les televisions ni en la productora amb les quals s'ha pogut contactar s'utilitzen els sistemes de síntesi de veu per generar productes audiovisuals. Tanmateix, l'opinió dels tècnics consultats és força variada. En un cas, s'indica que no interessen perquè el que agrada és la veu humana. Contràriament, s'exposa que es podrien utilitzar en programes automàtics que donin informació sobre la borsa o el temps i, també, en anuncis publicitaris, documentals i promocions pel gran estalvi econòmic que suposaria en la generació d'aquests productes. Aquesta última indicació s'ha extret de l'entrevista realitzada al representant tècnic del Departament d'Àudio de la productora de televisió Gestmusic. Tot i que alguns tècnics vegin viable aplicar veu sintètica per a diverses aplicacions, també indiquen que els sistemes de síntesi de veu haurien de madurar a nivell de naturalitat per poder produir diverses entonacions (veus agudes, greus, juvenils, serioses,...).
- Només dos dels departaments tècnics de l'últim grup (estudis de so, de doblatge i de postproducció) han utilitzat algun cop un sintetitzador de veu, però només per crear efectes en l'àmbit musical o per manipular veus. L'opinió general respecte a la implantació d'aquests sistemes de comunicació en un futur és molt semblant en tots els estudis consultats. La gran majoria de les persones entrevistades destaca que fins que els sistemes de síntesi de veu no estiguin més perfeccionats –en el sentit d'augmentar la naturalitat de la veu sintètica generada per tal de transmetre emocions de forma realista, tal com ho fa una persona–, la veu sintètica no es podrà utilitzar ni en el sector de la televisió ni en el de la ràdio.

Com a valoració global de la idea de la introducció dels sistemes de síntesi de veu en els mitjans de comunicació audiovisual, es pot dir que les opinions dels 19 tècnics entrevistats, en principi contraris a integrar-los en el procés de creació de continguts audiovisuals, podrien canviar si s'arribessin a sintetitzar de forma natural les emocions en la veu i s'aconseguissin veus sintètiques menys robòtiques i, per tant, més semblants a la veu natural produïda per l'ésser humà.

3.2. Usuaris potencials

Quant a les entrevistes realitzades en el context de les tecnologies per a les persones amb discapacitat visual, les entrevistes s'han realitzat a dos perfils diferents: 1) els tècnics que treballen en els mateixos mitjans de comunicació recollits a l'apartat anterior, per tal de conèixer la seva opinió respecte a l'ús de veu sintètica per a l'audiodescripció (tecnologia que ells ja coneixen), i 2) el sector de la població que pateix algun tipus de discapacitat visual, ja que és essencial considerar la seva opinió per tal de conèixer la viabilitat de la introducció de veu artificial en aquests medis.

La majoria de persones dedicades a les tecnologies del so (englobant-hi els tècnics de la ràdio, de la televisió i d'estudis d'àudio, de doblatge i de postproducció entrevistats) creu que es podria aplicar veu sintètica en l'audiodescripció si fos més natural i "creïble", tot i que, en alguns casos, es pensa que

tampoc no suposa un gran estalvi de temps i que no val la pena substituir la veu natural. Concretament, s'han recollit opinions en el sentit que els sistemes de síntesi de veu haurien de millorar molt quant a qualitat sintètica i, fins i tot, s'afirma que és més ràpid enregistrar-ho amb una persona. Tanmateix, en el conjunt de les entrevistes, n'hi ha hagut dues que destaquen especialment pel fet que són clarament diferents de les altres. Concretament, s'hi indica que:

- Abans d'incorporar les tecnologies de síntesi de veu a la producció audiovisual, s'hauria de preguntar a les persones amb discapacitat visual, que realment en són els usuaris finals, sobre la viabilitat d'usar veu sintètica per a l'audiodescripció i si no els agrada, caldria deixar de banda aquesta opció.
- Sempre és millor si l'entonació i la naturalitat del missatge són bones, però els costos poden ser un factor clau. En aquest sentit, tot i que la veu sintètica no sigui del tot natural, pot permetre abaratir els costos de la creació de l'àudio i, per tant, pot ser més rendible que contractar un locutor o locutora.

3.3. Usuaris amb discapacitat visual

Mitjançant la col·laboració amb l'ONCE, s'han pogut entrevistar 51 persones amb discapacitat visual de tot el territori espanyol. La distribució per territoris de les persones consultades és la següent: 16 persones de Madrid, 8 d'Andalusia, 5 de la Comunitat Valenciana, 3 de Catalunya, 3 de Galícia, 3 de les Illes Balears, 3 d'Astúries, 2 de Canàries i un sol representant a la resta de territoris (Cantàbria, País Basc, Castella i Lleó, etc.). La distribució per professions és la següent: 10 jubilats, 9 agents venedors de cupons, 7 fisioterapeutes, 2 estudiants, 2 mestres i un únic representant a la resta de professions (programador, músic, logopeda, periodista, advocat, administratiu, telefonista, treballador de banca, etc.). Per tant, en les entrevistes s'ha volgut recollir un ampli espectre de perfils d'usuaris potencials amb discapacitat visual (vegeu Torrens 2010 per a una enumeració detallada).

De les entrevistes realitzades a les persones amb discapacitat visual, ja sigui total o parcial, se n'extreuen dues idees força interessants relacionades amb els mitjans de comunicació:

- Gairebé totes les persones que han col·laborat responent al qüestionari creuen que en un futur es podria utilitzar veu sintètica per a l'audiodescripció a la televisió i al cinema. Indiquen que seria molt interessant que una veu els expliqués tot allò que no poden veure en programes de televisió, documentals, pel·lícules... Tot el que els permeti una normalització i una integració en el consum de productes audiovisuals és benvingut.
- Respecte a la introducció dels sistemes de síntesi de veu a la ràdio, les opinions són diverses. Més de la meitat creuen que és innecessari i prefereixen la veu humana. De la resta d'entrevistes, algunes consideren que pot ser útil, depenent de la qualitat de les veus sintètiques, i d'altres, tot i que ho accepten, no creuen que sigui imprescindible.

Finalment, es pot concloure que el dia que s'aconsegueixi naturalitat i emotivitat en les veus sintètiques, l'audiodescripció pot ser una bona via per introduir de forma progressiva els sistemes de síntesi de veu en el món de les produccions audiovisuals, ja que gairebé totes les persones amb discapacitat visual utilitzen aquests sistemes. Mentre s'espera aquest avenç en les veus, la viabilitat d'introduir els sistemes de síntesi de veu a la ràdio o a la televisió sembla difícil, però existeix l'opció d'utilitzar-los en sectors o en aplicacions en què no calgui l'expressivitat o bé es vulgui modelar una veu robòtica.

4. Adaptació del sistema de síntesi de La Salle al català

En l'àmbit tècnic del projecte, una de les fases clau ha estat l'encarregada de desenvolupar els recursos lingüístics i de processament del senyal per a la creació de les veus en català. El recursos lingüístics, com ara el sistema de transcripció fonètica, l'analitzador morfosintàctic, etc., que formen part del mòdul de processament del llenguatge natural (PLN) del sistema de síntesi són propis i han estat desenvolupats dins del marc del grup de recerca durant els darrers anys d'investigació. En canvi, les bases de dades de síntesi de veu en català són públiques i han estat desenvolupades pel grup de recerca TALP de la Universitat Politècnica de Catalunya, amb finançament de la Generalitat de Catalunya, en el marc del projecte FestCat (*http://gps-tsc.upc.es/veu/festcat*).

D'aquest projecte se n'han escollit les dues veus –Ona i Pau– que tenen més extensió, atès que el sistema de síntesi de veu del Grup de Recerca en Tecnologies Mèdia de La Salle (URL) està basat en la tècnica de selecció d'unitats en funció dels paràmetres predits pel model prosòdic.

Un cop es disposa dels fitxers de veu, cal "crear una nova veu" pel sistema de síntesi, és a dir, cal processar les mostres de veu per tal que siguin útils per generar veu sintètica. La creació d'una nova veu consta de tres parts principals:

1. La segmentació de la base de dades en unitats de síntesi, que s'encarreguen de determinar l'inici i el final de cada una de les unitats acústiques (difonemes, en aquest cas) que integren els missatges enregistrats en els fitxers de veu.
2. La indexació i la parametrització de la base de dades, que s'encarreguen de generar el conjunt de fitxers en format XML que contenen els paràmetres que descriuen el contingut acústic de la base de dades (durada, energia, freqüència fonamental de les unitats). Alhora, cal ajustar la funció de cost de selecció, qüestió que implica, per una banda, precalcular tots els costos de les unitats de la base de dades i, per l'altra, ajustar els pesos de la funció de cost (Alías *et al.* 2011).
3. L'entrenament del model prosòdic, que és l'encarregat de determinar la pronúncia més adequada d'un text d'entrada a sintetitzar a partir de l'extracció de patrons prosòdics que s'extreuen de les mostres de veu disponibles (Iriondo *et al.* 2007).

Un cop finalitzades aquestes tres fases, ja es disposa de les veus Ona i Pau integrades en el sistema de síntesi de veu de La Salle per tal de realitzar els experiments que tenen l'objectiu d'analitzar la viabilitat de l'ús de la síntesi de veu en produccions audiovisuals i que es descriuen tot seguit.

5. Experiments i resultats

En l'àmbit de la síntesi de la parla es poden avaluar diferents característiques, com ara la intel·ligibilitat, la naturalitat i l'expressivitat. En algunes aplicacions, com per exemple, en les màquines parlants per a persones invidents, la intel·ligibilitat de la parla a alta velocitat és més important que la naturalitat (Llisterri *et al.* 1993). En canvi, una prosòdia correcta i una naturalitat elevada són essencials en la majoria d'aplicacions multimèdia. L'avaluació es pot realitzar a diferents nivells (segment, paraula, frase o paràgraf) i amb diferents tipus de proves (Campbell 2007).

Amb la finalitat de disposar d'una avaluació subjectiva de la viabilitat de l'ús de la síntesi de veu a l'hora de generar material audiovisual, s'han preparat dos tests perceptius: un d'anuncis publicitaris i un altre de notícies. Per cada test, es van preparar un conjunt de parelles d'estímuls. Cada parella tenia el mateix contingut verbal però una estava generada amb el sistema de síntesi i l'altra era llegida per una persona. Un cop preparats els estímuls, es va decidir el tipus de prova més adequada per presentar-los als oients i la metodologia d'avaluació d'aquests estímuls. En el cas dels anuncis, només portaven el canal d'àudio, mentre que en el cas de les notícies eren vídeos amb imatges relacionades amb la notícia i el canal d'àudio format per la pista de so de fons (música, soroll de carrer, veus, etc.) superposada a la pista de veu en *off*.

Com ja s'ha assenyalat, l'objectiu de la prova ha consistit a avaluar la síntesi de la parla en anuncis o en notícies. Es disposava d'una parella de fitxers d'àudio (anuncis) o de vídeo (notícies) per cada element que s'havia d'avaluar. Es van plantejar diferents possibilitats de presentació dels estímuls (de manera individual o per parelles) i d'escales de puntuació. A partir de la recomanació P.800 de la Unió Internacional de Telecomunicacions (UIT) (UIT-T 1996), es va escollir l'índex d'avaluació comparativa *Comparison Mean Opinion Score* (CMOS), que permet comparar dos estímuls, *A* i *B*, com:

- A molt millor que B
- A millor que B
- A lleugerament millor que B
- Cap preferència
- B lleugerament millor que A
- B millor que A
- B molt millor que A

Amb aquesta escala, els oients van poder avaluar comparativament els dos estímuls presentats escoltant-los tants cops com calia.

5.1. Anuncis publicitaris

Per avaluar l'ús de la síntesi de la parla en situacions reals, es va elaborar un test amb set anuncis publicitaris. Per cada anunci, es van generar dos fitxers de so, un a partir de la lectura de l'anunci per part d'una locutora *amateur* i l'altre, utilitzant el nostre sintetitzador de parla en català.

El test es va realitzar amb la plataforma en línia TRUE (Testing platfoRm for mUltimedia Evaluation) (Planet *et al*. 2008), que permet dissenyar i realitzar el test de forma remota.

Per cada parella de àudios associats al mateix anunci, al participant del test se li van formular dues preguntes:

1. "Els àudios següents (*A* el de dalt, *B* el de baix) corresponen a dues lectures d'anuncis publicitaris. No es tracta d'avaluar si t'agrada més la veu d'una dona o de l'altra, sinó, per a un ús en publicitat, indica la teva preferència, fixant-te en la NATURALITAT de la pronúncia i de l'entonació:"
2. "Quant a la INTEL·LIGIBILITAT, què et sembla?"

El test el van realitzar 25 oients (12 dones i 13 homes) d'edats compreses entre els 18 i els 66 anys. Els resultats de preferència obtinguts amb aquest test es mostren a la figura 2, on *A* representa la veu natural i *B*, la veu generada amb el sintetitzador. Els resultats, com és d'esperar, mostren una preferència clara per la veu natural especialment quant a naturalitat, tot i que en intel·ligibilitat la diferència no és tan gran.

5.2. Vídeos de notícies

En aquest experiment s'ha volgut afegir a la veu dos components habituals en el material audiovisual: la imatge i una pista de so addicional a la de veu. Es va preparar un test amb tres parelles de notícies. A partir de material extret de YouTube i de la veu generada amb el nostre sintetitzador, es van generar vídeos de notícies que contenien tres pistes: la de vídeo pròpiament i dues d'àudio (so de fons i veu).

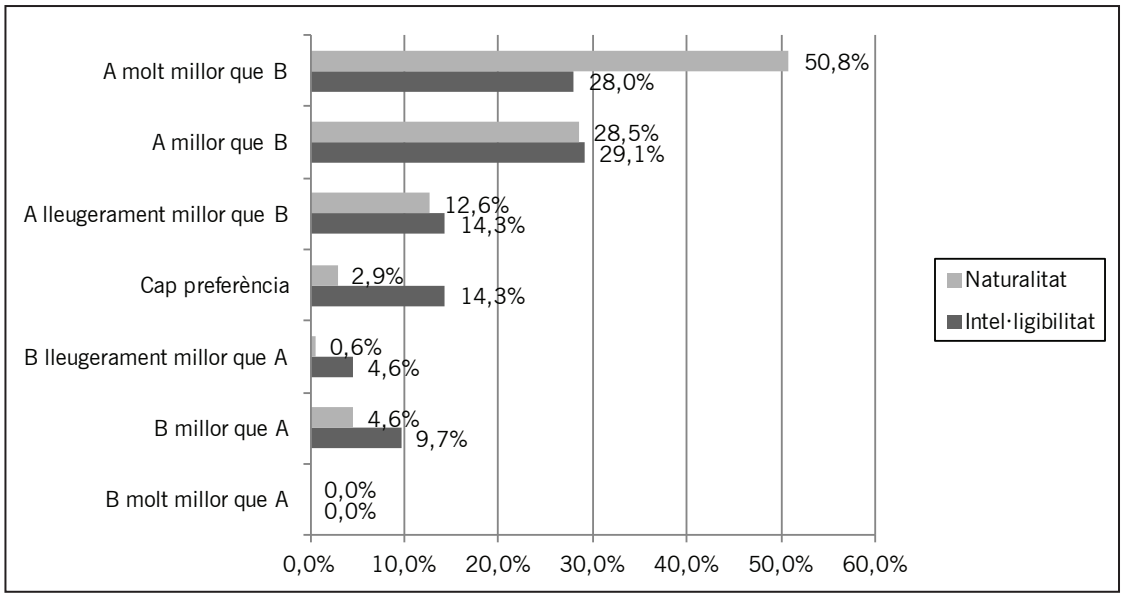
El test també es va realitzar amb la plataforma TRUE i es tractava d'un CMOS de set categories. Hi van participar 20 persones (17 homes i 3 dones) d'edats compreses entre els 24 i els 41 anys. Als usuaris no se'ls va informar de l'origen de les dues veus. Al final del test es va preguntar el sexe i l'edat del participant i si era expert en tecnologies de la parla, i se li va formular dues preguntes de resposta oberta:

1. "La veu del vídeo de sota ha estat generada per ordinador, què t'ha semblat?"
2. "Creus que és factible l'ús de síntesi de veu per explicar notícies en programes que es generin automàticament?"

Els resultats obtinguts es mostren a la figura 3, on es pot observar com la resposta majoritària és que la veu natural és lleugerament millor que la sintètica (46,3%). És important destacar que pràcticament un 26% de les respostes (18,5 % de *cap preferència* més un 7,4% de *la veu sintètica és lleugerament millor que la natural*) indiquen que la veu sintètica és acceptable en aquest context.

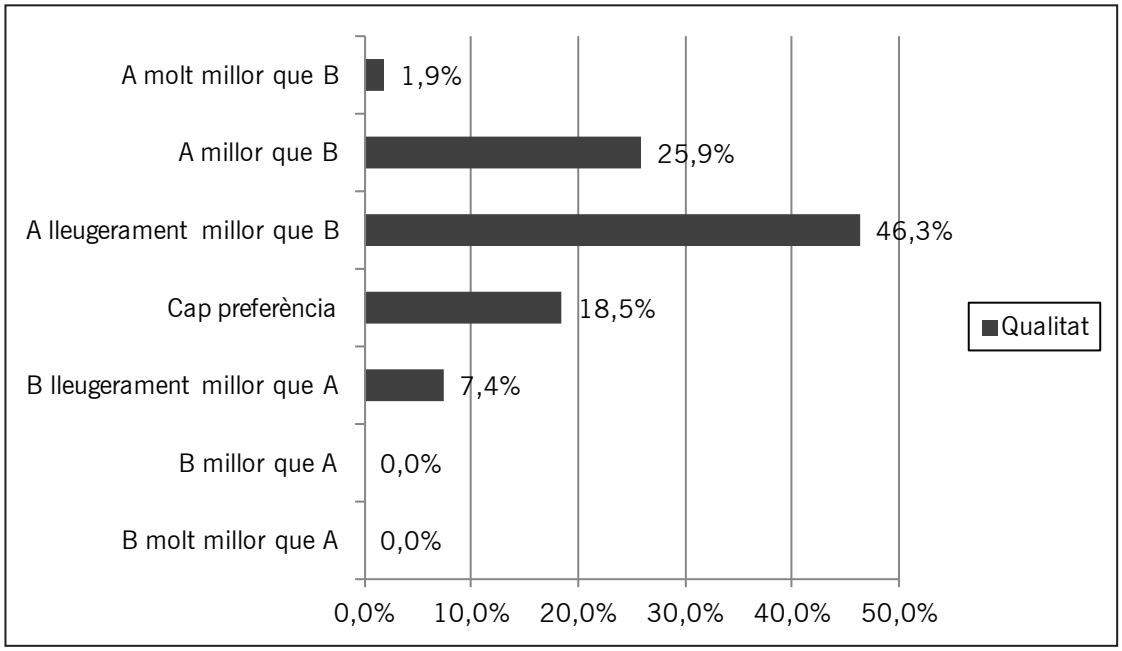
Si analitzem les respostes dels participants on han manifestat,

Figura 2. Resultats del test d'anuncis publicitaris quant a intel·ligibilitat i naturalitat. A es correspon a la veu natural i B, a la veu sintetitzada



Font: Elaboració pròpia.

Figura 3. Resultats del test de vídeos de notícies quant a qualitat de la veu en off. A es correspon a la veu natural i B, a la veu sintetitzada



Font: Elaboració pròpia.

després de fer el test, la seva opinió respecte a l'ús de la síntesi de la parla per generar notícies, podem destacar dues idees generals. En primer lloc, que els oients són molt sensibles a errors puntuals en una determinada part del text i que falta millorar l'expressivitat i el ritme. En segon lloc, l'opinió majoritària és que l'ús d'aquesta tecnologia el veuen factible per generar notícies d'última hora, per exemple per a la web o en programes de generació semiautomàtica. Per il·lustrar aquestes dues conclusions, a continuació es reprodueix un conjunt ampli de les respostes obtingudes en les taules 1 i 2.

Taula 1. Selecció de respostes a la pregunta "La veu del vídeo de sota ha estat generada per ordinador, què t'ha semblat?"

"Bastant acceptable, encara que una mica lenta i amb algunes errades en sons concrets."
"Bona qualitat en general, tot i que hi ha algunes discontinuïtats i salts en l'entonació."
"Bastant aconseguida, però en certs moments es nota que no és humana."
"De vegades molt bé (millor que l'original i tot), d'altres no. Els "galls" puntuals en fan baixar la qualitat global."
"Poc natural, tot i que s'hi notava una mica d'expressivitat i la qualitat de l'àudio estava força ben aconseguida. Potser problemes en el fet de mantenir un ritme constant, es noten salts de ritme."
"La veu és una mica metàl·lica. L'entonació no és prou natural. En tot moment notes, sens dubte, que t'està parlant una màquina. Malgrat tot, el missatge s'entén correctament."
"Bastant bé, sobretot en el primer. El soroll de fons dissimula els errors. En funció de la temàtica, l'estil de locució hauria de variar (p. ex., en ambient festiu, parla més àgil)."
"Es nota que és una veu sintètica però no és molest perquè s'integra bé amb la música i les imatges, i la seva qualitat permet que s'entengui bé tot el que diu, fins i tot millor, de vegades, que la real."
"Força bona quant a versemblança de veu humana i d'entonació. El fet que la converteix de menys qualitat que la humana són uns sorolls, "clics", que apareixen de tant en tant."
"En el primer test la qualitat era prou bona, mentre que en la resta la qualitat ha decaigut. Es nota bastant la concatenació entre unitats."
"Prou bona; el principal problema són els artefactes de coarticulació, que resten naturalitat a la veu."
"Bastant bona tenint en compte que és àudio sintètic. De tota manera, es nota bastant que no és una veu humana natural."
"Qualitat acceptable. L'únic problema que detecto que es repeteix sovint és l'allargament/arrossegament d'algunes vocals i consonants."
"La veu és correcta i clara, però de tant en tant fa sons estranys i sona com distorsionada."

Font: Elaboració pròpia.

Si comparem els resultats amb el test d'anuncis publicitaris podem comprovar que el fet d'afegir-hi vídeo i so de fons ajuda a dissimular els errors de síntesi i a desviar l'atenció, amb la qual cosa millora l'acceptabilitat d'utilitzar veu sintètica.

Els fitxers d'àudio i de vídeo generats pels experiments es poden trobar al lloc web següent: <http://www.salle.url.edu/portal/departaments/home-depts-DTM-projectes-info?id_projecte=67>

6. Conclusions i línies de futur

En aquest treball, després de revisar l'estat de la qüestió en l'àmbit de la síntesi de veu (també conegut com a *sistemes de conversió de text en parla*), s'ha estudiat la situació d'aquesta tecnologia a Catalunya i, concretament, en l'àmbit de les produccions audiovisuals. En l'actualitat hi ha diversos centres de recerca i empreses que treballen en el desenvolupament i la millora dels sistemes de síntesi de la parla en català. Tanmateix, la implantació d'aquests sistemes en el context de la generació de produccions audiovisuals encara és molt reduïda. Atesa aquesta situació, s'ha avaluat la viabilitat de la implantació d'aquesta tecnologia en el món de les produccions audiovisuals, a partir d'un treball de camp que ha consistit en diverses entrevistes tant a personal tècnic com a usuaris potencials, així com un conjunt d'experiments dissenyats per estudiar el grau d'acceptació de la síntesi en exemples reals.

Tant de les entrevistes com dels experiments realitzats, es pot concloure que l'ús de veu sintètica en contingut *broadcast* pot ser una realitat en els propers anys si es milloren certs aspectes relacionats amb el fet d'aconseguir l'expressivitat pròpia del contingut. Un altre aspecte important és el nombre de modes que formen part del contingut. Si la veu va acompanyada d'uns altres elements d'àudio superposats així com del canal de vídeo, llavors l'ús de veu sintètica es preveu més factible. En

Taula 2. Selecció de respostes a la pregunta "Creus que és factible l'ús de síntesi de veu per explicar notícies en programes que es generin automàticament?"

"Sí, ho veig factible i interessant."
"Sí, especialment si es tracta de notícies curtes i de darrera hora, de forma que sigui més adequada una producció semiautomatitzada que faci possible disposar amb més celeritat dels continguts."
"En un futur ha de ser més que viable."
"No seria factible per a un telenotícies per a televisió, per exemple, però potser sí per a contingut al web, on la qualitat del contingut no és el que prima, sinó el contingut en si mateix."
"Li falta naturalitat i expressivitat, els quals ajuden a fer una notícia més atractiva. No obstant això, la intel·ligibilitat és molt bona i el missatge es pot transmetre perfectament. Seria factible."
"Sí. Tot i la falta de naturalitat, que és millorable. El resultat és prou satisfactori."
"Sí. Els petits problemes amb la síntesi queden sota la pista sonora de la notícia i no suposen un problema per entendre-la. A més, formalment la locució és correcta (to neutre)."
"Si. És igual d'intel·ligible que la veu humana."
"Sí, però depenent de l'àmbit en què s'apliqui. Si és en plataformes web, crec que a nivell d'usuari es pot acceptar aquesta qualitat."
"Sí, sempre que s'evitin els artefactes esmentats més amunt."
"Sí que em sembla factible, però no tal com està ara el TTS. Encara li falta més naturalitat. La veu que genera ara resulta massa desagradable per a un locutor al qual has d'escoltar habitualment."
"La comprensió és perfecta. Si es pogués millorar el tema de les petites distorsions faria el seguiment de les notícies més agradable."

Font: Elaboració pròpia.

canvi, en continguts on només hi ha veu (p. ex. un anunci publicitari per a ràdio), l'exigència dels oients sobre la qualitat d'aquesta veu és molt més gran.

L'ús de la síntesi de veu (no només en català) com a mitjà de suport per disposar de sistemes més automatitzats i capaços de servir continguts en un format més natural i que permeti també més capacitat d'incloure tothom, és un dels reptes en què ja s'està treballant. En aquest context, hi ha estudis *(http://www.daisy.org/benefits-text-speech-technology-and-audio-visual-presentation)* que afirmen que la inclusió de la modalitat acústica com a forma alternativa de presentar continguts purament en un format textual, per exemple, permeten augmentar la capacitat de retenció, essent per tant una forma apropiada de presentació per a activitats d'aprenentatge en entorns d'un caire més audiovisual. Existeixen ja empreses que basen el seu negoci a donar serveis de veu automatitzada a partir d'informació textual, com IVO Software *(http://www.ivona.com)*, Odiogo *(http://www.odiogo.com/)* o NextUp.com *(http://www.nextup.com)*, que permeten, per exemple, incorporar informació oral a un web o donar solucions per generar veu de forma automàtica a partir de documents de text. Tot i que solucions com aquestes ens permetran cada cop més disposar de sistemes amb un grau més elevat d'adaptació a la persona usuària, encara estem lluny de veure sistemes que actuïn com ho fem les persones i evitin qualsevol mínim artefacte sonor o accentuïn els rangs d'expressivitat propis d'una veu humana. En tot cas, les solucions que avui dia podem trobar arreu són solucions que encara no ens permeten trobar un missatge de qualitat equiparable a una locució parlada per una persona real en una conversa real, però ens hi anem apropant, i els nous paradigmes d'interacció i d'intercanvi amb els proveïdors de continguts que el futur ens depara ben segur que tindran en compte l'ús de la tecnologia de la síntesi de veu com a eina molt vàlida per emfatitzar o redundar en un missatge més proper a l'humà.

Per tal de possibilitar la utilització de la síntesi de la parla en continguts audiovisuals cal seguir avançant en les línies d'investigació següents:

- Millorar l'expressivitat de la parla generada per adaptar els trets suprasegmentals (ritme, entonació, intensitat, èmfasi, etc.) a les característiques pròpies del mode de locució de cada tipus de contingut. Aquesta millora es pot aconseguir si es compta amb l'aportació dels coneixements d'experts en el camp de la comunicació audiovisual.
- Millorar la qualitat segmental de la síntesi per evitar arte-

factes sonors, ja que cal tenir en compte que l'oïda humana és molt sensible a aquests petits errors. En aquest aspecte, hi influeixen errors relacionats amb la fonètica i amb el processament del senyal. Per tant, seria desitjable comptar amb experts en fonètica que aportessin coneixement per millorar, per exemple, les regles de transcripció fonètica, especialment les que fan referència a la coarticulació. Quant al processament del senyal, hi ha camí a recórrer en la parametrització i el modelat de la veu per poder dur a terme modificacions de les seves característiques sense distorsionar-la.

- Aconseguir nous mètodes per generar noves veus mitjançant tècniques de transformació de veu que permetin augmentar el nombre de veus d'alta qualitat disponibles en un idioma determinat.

Referències

Alías, F.; Formiga, L.; Llorà, X. "Efficient and reliable perceptual weight tuning for unit-selection Text-to-Speech synthesis based on active interactive genetic algorithms: a proof-of-concept". *Speech Communication,* vol. 53 (5), p. 786-800, maig-juny, 2011.

Campbell, N. "Evaluation of Text and Speech Systems". *Text, Speech and Language Technology*, vol. 37 (2007), p. 29-64, Springer, Dordrecht.

Camps, J.; Bailly, G.; Martí, J. "Synthèse à partir du texte pour le catalan". *Proceedings of 19èmes Journeés d'Études sur la Parole* (1992), p. 329–333, Brussel·les, Bèlgica.

Guaus, R.; Iriondo, I. "Diphone based Unit Selection for Catalan Text-to-Speech Synthesis". *Proceedings of Workshop on Text, Speech and Dialogue* (2000), Brno, República Txeca.

Iriondo,I.; Alías, F.; Melenchón, J.; Llorca, M. A. "Modeling and Synthesizing Emotional Speech for Catalan Text-to-Speech Synthesis". *Tutorial and Research Workshop on Affective Dialog Systems, Lecture Notes in Artificial Intelligence*, núm. 3068 (2004), Springer Verlag, p. 197-208, Kloster Irsee, Alemanya.

Iriondo,I.; Socoró, J.C.; Alías, F. "Prosody Modelling of Spanish for Expressive Speech Synthesis". *International Conference on Acoustics, Speech and Signal Processing*, vol. IV (2007), p. 821-824, Hawaii, Estats Units.

Llisterri, J.; Fernández, N.; Gudayol, F.; Poyatos, J. J.; Martí, J. "Testing user's acceptance of Ciber232, a text to speech system used by blind persons". *Proceedings of the ESCA Workshop on Speech and Language Technology for Disabled Persons* (1993), p. 203–206, Estocolm, Suècia.

Planet, S.; Iriondo, I.; Martínez, E.; Montero, J.A. "TRUE: an online testing platform for multimedia evaluation". Proceedings of the Second *International Workshop on Emotion: Corpora for Research on Emotion and Affect at the 6th Conference on Language Resources & Evaluation* (2008), Marràqueix, Marroc.

Rodríguez, M.A.; Escalada, J. G.; Armenta, A.; Garrido, J.M. "Nuevo módulo de análisis prosódico del conversor texto-voz multilingüe de Telefónica I+D". *Actas de las V Jornadas en Tecnología del Habla* (2008), p. 157-160.

Torrens, A."Estudi sobre la utilització de les tecnologies de síntesi de veu en els mitjans audiovisuals de Catalunya". *Treball final de carrera* (2010). Barcelona: La Salle - Universitat Ramon Llull.

UIT-T (1996). "Recomendación P.800: Métodos de determinación subjetiva de la calidad de transmisión*". Sector de Normalización de las Telecomunicaciones de Unión Internacional de Telecomunicaciones*.
<http://www.itu.int/rec/T-REC-P.800-199608-I/es>