ELS TALPS TAMBÉ PARLEN.

Línies de recerca en síntesi de la parla al centre TALP

Ignasi Esquerra, Jordi Adell, Pablo D. Agüero, Antonio Bonafonte,
Helenca Duxans, Asunción Moreno, Javier Pérez, David Sündermann

Centre de Tecnologies i Aplicacions del Llenguatge i la Parla
Departament de Teoria del Senyal i Comunicacions
Universitat Politècnica de Catalunya
(www.talp.upc.es)

El Centre de Tecnologies i Aplicacions del Llenguatge i la Parla (TALP) és un centre de recerca interdepartamental de la Universitat Politècnica de Catalunya, l'àmbit tecnològic del qual és el tractament automàtic del llenguatge natural, tant en la modalitat oral com en l'escrita, amb l'objectiu d'ajudar a superar les barreres lingüístiques i millorar l'accessibilitat dels sistemes d'informació.

Hi treballen al voltant d'una quarantena d'investigadors, la majoria professors de les titulacions de Telecomunicació i Informàtica de la UPC, i està format pel grup de Tractament de la Parla (departament de Teoria de Senyal i Comunicacions) i el grup de Tractament del Llenguatge Natural (departament de Llenguatges i Sistemes Informàtics).

El centre TALP pertany a la xarxa ELSNET (European Network of Excellence in Human Language Technologies), i els seus dos grups són grups de recerca consolidats i membres del Centre de Referència en Enginyeria Lingüística (CREL) de la Generalitat de Catalunya.

El TALP, com a centre de R+D, fomenta la transferència de coneixements, experiència i tecnologia, especialment per mitjà de la cooperació amb institucions públiques i empreses en projectes de recerca aplicada i desenvolupament. Actualment participa activament en diversos projectes d'àmbit nacional, europeu i internacional: CHIL (Computers in the Human Interaction Loop), TC-STAR (Technology and Corporafor Speech to Speech Translation), HOPS (Enabling an Intelligent Natural Language Based Hub for the Deployment of Advanced Semantically Enriched Multi-channel Mass-scale Online Public Services), SIMILAR (Human-machine interface similar to human-human communication), BIOSEC (Biometrics & Security), ALIADO (Tecnologías del habla y el lenguaje para un asistente personal), MEANING (Developing Multilingual Web-scale Language Technologies).

1. La síntesi de parla

Una de les àrees principals de treball en el Centre de Tecnologies i Aplicacions del Llenguatge i la Parla (TALP) és la conversió text-parla, és a dir, la generació de veu sintètica a partir de text. Aquests sistemes han d'analitzar els texts que es volen llegir (tractament del text), convertir-los en una representació dels sons que cal pronunciar i com s'han d'expressar (transcripció fonètica i prosòdica), i finalment produir acústicament aquests sons (generació del senyal). Habitualment aquests sistemes es coneixen amb el nom de sistemes TTS (Text-to-Speech).

La majoria de sintetitzadors actuals no generen la veu de forma completament articial sinó que ho fan mitjançant la manipulació i encadenament de segments de veu natural prèviament enregistrada, extrets de grans bases de dades orals. Gràcies a tècniques de processament de senyal i d'aprenentatge automàtic es poden aconseguir models de producció de la veu i de generació de la prosòdia (entonació, durada, intensitat...) que fan la veu sintètica molt més natural i intel·ligible.

Les llengües amb les que treballem en el grup de síntesi del TALP són el català, l'espanyol i, més recentment, també l'anglès. Alguns dels mòduls del sistema són comuns a totes les llengües; d'altres tenen una forta dependència de l'idioma i no sempre es poden aprofitar els resultats obtinguts per altres idiomes.

Per tal de millorar la qualitat del nostre sistema, en el grup de síntesi del TALP estem treballant en diverses línies de recerca. D'una banda, el desenvolupament de noves veus per al sintetitzador és un procés que consumeix molt temps, degut a la revisió manual de la segmentació. Actualment s'està investigant en la generació de marques a partir de models ocults de Markov i característiques acústiques, que són després corregides per un sistema basat en característiques fonètiques de forma automàtica. Així mateix, els models de prosòdia són molt importants per tal de generar una síntesi que soni natural. L'anàlisi i parametrització de diverses bases de dades en els tres idiomes, i l'obtenció de models de prosòdia és un altre dels temes de treball. En quant al mòdul de síntesi, la bona concatenació de les unitats és bàsica per tal que la parla generada soni fluida i natural. Per això calen mètodes de tractament del senyal com els models sinusoidals o síntesi PSOLA que estem programant per al sistema.

D'altra banda, els sistemes de conversió text-parla cada vegada més s'apliquen en àmbits més diversos, i es no poden limitar a generar veu sintètica amb unes caracaterístiques d'entonació, velocitat o timbre de veu sempre iguals. Per exemple, en aplicacions de traducció veu-veu o en dispositius per a persones discapacitades, és desitjable que el sistema generi una veu personalitzada, que soni natural i espontània. Mitjançant tècniques de conversió de veu es pot transformar les característiques d'una veu i convertir-la en un nou locutor. Igualment, s'està analitzant bases de dades de parla espontània i emocional, per tal de desenvolupar models per a la síntesi i generar veus molt més variades.

2. Segmentació de les bases de dades

A l'hora de crear un sintetitzador basat en concatenació d'unitats cal, abans de res, tenir una base de dades de veu correctament dissenyada i etiquetada. És necessari que el disseny de la base de dades asseguri la representativitat de la variabilitat lingüística de la llengua que volem sintetitzar; és a dir, que cal que apareguin el màxim nombre d'unitats diferents en contexts diferents, i possiblement voldrem que les que apareixen més a la llengua també apareguin més a la base de dades. També cal tenir en compte que la base de dades ha de ser enregistrada amb

la màxima cura possible, en el sentit d'eliminar quantes més fonts de soroll millor i mantenir les condicions d'enregistrament durant totes les sessions.

Un cop la base de dades està ben dissenyada i s'ha enregistrat amb cura, llavors comença el treball d'etiquetar-la i dotar-la de la informació necessària per a la cerca i selecció de les unitats. En aquest sentit, cal donar nom a la unitats, és a dir, cal fer la transcripció fonètica del text i cal trobar els extrems de cada unitat, és a dir, la segmentació de veu. D'altra banda, habitualment la concatenació es fa de forma síncrona amb el període fonamental de la veu, i per això cal també marcar els instants d'inici del pols glotal.

Aquests processos que hem enumerat es poden realitzar de forma manual. En aquest cas, una o vàries persones es dediquen a escoltar i a mirar el senyal de veu i manualment van posant les marques de segmentació fent la transcripció fonètica i posant també les marques de pols glotal. Més habitualment es fan servir mètodes automàtics i posteriorment persones especialistes supervisen el resultat d'aquests mètodes corregint la posició de les marques i ajustant la transcripció fonètica a la realment pronunciada.

El desig dels investigadors en síntesi de veu és el d'arribar a tenir mètodes totalment automàtics que puguin arribar a obtenir aquesta informació. Nosaltres també estem treballant en aquest sentit. La transcripció fonètica es pot fer mitjançant un diccionari o regles, fins i tot es poden combinar els dos mètodes, els diccionaris per la transcripció canònica i les regles per a trobar les variacions fonètiques (al·lòfons) que correspongui en cada cas. Finalment, la tasca més costosa de realitzar manualment és la segmentació. Existeixen una gran varietat de tècniques, les més utilitzades són les basades en Models Ocults de Markov (HMM) o alineament per Programació Dinàmica, encara que també n'hi ha d'altres que es basen en Xarxes Neuronals o mescles de Gaussianes.

Una de les línies actuals de recerca en aquest camp és el desenvolupament d'un mètode per a la segmentació basat en la correcció específica de les fronteres generades pels HMM fent servir característiques fonètiques. Aquest mètode fa servir un arbre de regressió per determinar el desplaçament de cada frontera calculat a partir d'un etiquetatge manual d'una petita part del corpus. Aquest mètode, tot i donar un molt bon resultat en avaluacions objectives, no aconsegueix superar la valoració perceptual aconseguida pels models ocults de Markov, que aconsegueixen un resultat semblant al de la segmentació supervisada manualment.

3. Models de prosòdia

Per a que la veu sintetitzada presenti una bona qualitat i sigui avaluada com a bona per part dels qui l'escolten, és necessari que, a banda d'una bona intel·ligibilitat (que s'aconsegueix bàsicament per una bona elecció i concatenació de les unitats fonètiques), també tingui una bona entonació, velocitat de pronunciació o evoluació de la intensitat de la veu. Aquest conjunt de característiques és el que s'engloba normalment sota el terme de prosòdia.

L'entonació està relacionada amb el contorn de freqüència fonamental, produida per la variació del període d'oscil·lació de les cordes vocals, i transmet informació lingüística i paralingüística, com per exemple de si es tracta d'una afirmació o pregunta, o si el locutor és un adult o un infant.

Els treballs de recerca sobre entonació tenen l'objectiu de desenvolupar models generats automàticament a partir de corpus. Disposem de corpus en espanyol (SpeechDat, Interface), en català i en anglès (Boston University Radio Speech Corpus). Tots els corpus estan

convenientment segmentats fonèticament, i etiquetats amb marques d'accents, síl·labes i categories sintàctiques. Els patrons d'entonació es basen en models superposicionals: dues components relacionades amb el grup entonatiu i el grup accentual són sumades per a obtenir un contorn de freqüència fonamental. Actualment estem treballant amb dos tipus de models. El primer model està basat en el model de Fujisaki, l'extracció dels paràmetres del qual s'ha fet amb restriccions lingüístiques per tal d'obtenir una parametrització més consistent. El segon model està basat en una representació paramètrica usant corbes de Bézier.

Igualment, estem treballant en l'obtenció de models per a les durades dels fonemes i en models per a l'assignació i quantització de les pauses. Ambdos models fan servir tècniques de classificació mitjançant arbres regressió (CART).

4. Tècniques de síntesi

En la generació automàtica de veu, els paràmetres d'entrada al sistema són els fonemes (o unitats) que formen part del missatge, i les característiques prosòdiques de com cal pronunciar-los (durada de les unitats, entonació, ritme, freqüència, etc.).

De tècniques de generació de veu n'hi ha moltes, i podem dividir-les en dues grans famílies, en funció de si fan servir (o intenten emular) o no el model de producció humà de veu. Ara per ara, els sistemes que fan servir models de l'excitació glotal i del tracte vocal no aconsegueixen igualar els sistemes que es dediquen a reproduir segments de veu prèviament enregistrada.

Aquests darrers sistemes són els més emprats actualment. Les unitats que es concatenen, així com les tècniques utilitzades per a ajuntar-les per obtenir la forma d'ona final, depenen de cada sintetitzador. Al sistema de conversió text-parla desenvolupat en la UPC utilitzem el mètode TD-PSOLA que permet modificar les durades i freqüències de les unitats per a adaptar-les a les que determina del mòdul de generació prosòdica. Aquesta tècnica es basa en l'encavalcament directe dels trossos de senyal, convenientment enfinestrats (suavitzats) per tal de disminuïr les discontinuïtats en els punts de concatenació. Com a unitat elemental per a la síntesi, utilitzem els semifonemes amb context. Cadascun d'aquests semifonemes correspon a una de les dues parts resultants de dividir un fonema qualsevol en dos trossos no necessàriament iguals, el primera amb informació del context previ, i el segon amb informació del posterior. El semifonema té l'avantatge de permetre formar difonemes (unions de dos fonemes) i trifonemes (unió de tres), les unitats més usades en síntesi per concatenació, de forma senzilla.

Malauradament, la naturalitat i qualitat d'aquests sistemes és molt depenent de la quantitat i diversitat de les unitats de la base de dades, ja que en el fons, el sistema el que fa és repetir trossos de veu enregistrats. Com que no és possible tenir enregistrades totes les unitats que es necessitaran en un futur, cal emprar tècniques de transformació a les unitats existents per tant d'adequar-les a les unitats (diferents) requerides per la síntesi. És d'esperar que les modificacions siguin mínimes, ja que la tasca del mòdul de selecció d'unitats és seleccionar l'unitat de la base de dades més propera (segons una mesura de proximitat espectral). El sistema TTS actual incorpora diverses tècniques per tal d'aconseguir unitats òptimes per la concatenació, modificant la prosòdia de les unitats prèviament a la concatenació. Les modificacions que el sistema TD-PSOLA permet són bastant limitades, reduïnt-se a modificar la freqüencia fonamental i la durada de l'unitat a concatenar.

Per aquest motiu, estem treballant ara en la reparametrització de la veu emmagatzemada per tal d'incloure informació del model de producció humà en les unitats. D'aquesta manera

obtindríem unitats per a la concatenació molt més adients per a ser sotmeses a modificacions espectrals i prosòdiques d'alta qualitat. La línia més activa en aquests moments està enfocada al model de tracte vocal excitat amb un pols glotal. La tasca està en estimar, a partir dels enregistrmants de veu disponibles, els paràmetres del filtre i pols glotal que resulten en la veu resintetitzada més propera a l'original.

5. Conversió de veu

Un sistema de conversió de veu modifica la veu d'un locutor (anomenat locutor font) per a què es percebi com si estigués parlant un altre locutor conegut (anomenat locutor objectiu). Les aplicacions dels sistemes de conversió de veu es poden trobar en molts camps, per exemple en la personalització de sistemes de conversió text-parla, en la traducció automàtica mantenint les característiques del locutor en la generació de la veu en la nova llengua, en la creació d'eines pedagògiques per l'estudi de llengües estrangeres, en el camp mèdic com a ajuda per a millorar la veu de persones amb problemes de la parla o en el camp de l'oci (karaokes, doblatge de pel·lícules, etc.).

L'objectiu del treball en conversió de veu que es realitza en el centre TALP és desenvolupar un sistema de conversió de veu com a bloc posterior a un TTS, per tal de no haver de produir i emmagatzemar diferents bases de dades, una per a cada locutor. Molts del mètodes que s'han proposat per a la conversió de veu es basen en el model de pols glotal més tracte vocal per a la producció de la parla, aprenent una funció de mapeig entre les característiques de tracte vocal del locutor font i de l'objectiu i predint el senyal residual transformat a partir del tracte vocal. Una de les funcions de mapeig més utilitzades utilitza una mescla de Gaussianes (GMM) per a modelar l'espai acústic conjunt del locutor font i del locutor objectiu. Per estimar l'espai acústic conjunt es necessiten vectors font-objectiu alineats. Aquest tipus de funcions de mapeig treballen tram a tram, utilitzant només informació acústica per estimar la funció i convertir les veus.

La recerca més recent s'ha basat sobretot en dos punts: en incloure característiques dinámiques en el model acústic, extenent l'ús dels GMM cap a HMM (models ocults de Markov), i en la introducció d'informació fonètica a través d'arbres de decisió CART per a l'estimació de la funció de mapeig i la conversió de veus. Els resultats perceptuals confirmen que la inclusió d'informació fonética millora la conversió de veu.

6. Síntesi d'emocions

En els darrers anys la síntesi d'emocions ha sigut objecte de gran interès per part d'investigadors de tot el món. Inicialment els sintetitzadors es basaven en regles per a generar les diferents variants de prosòdia, tipus de veu o estils de parla, i com a sistema de síntesi utilitzaven models de formants. En canvi, els sintetitzadors actuals fan servir tècniques de selecció d'unitats a partir de corpus, amb els que s'aconsegueixen veus de major qualitat però que rarament permeten un gran control sobre la prosòdia o el tipus de fonació de la veu.

En el centre TALP estem realitzant diversos treballs amb l'objectiu d'aconseguir una conversió text-parla més versàtil i que soni més natural i espontània. Es disposa d'una base de dades oral enregistrada per dos locutors (masculí i femení) que conté frases curtes i paràgrafs simulant 6 emocions (alegria, fàstic, enfadat, por, sorpresa i tristesa) i 4 estils de parla (fluix, fort, lent i ràpid) a més de l'estil neutre.

En uns primers experiments es va realitzar una transformació de frases prosòdicament neutres a frases amb emoció mitjançant tècniques de re-síntesi PSOLA. Aquest procés consisteix en analitzar els patrons d'entonació i durada de les unitats fonètiques de la frase amb emoció, i transportar-los sobre la mateixa frase expressada de forma neutra, de tal manera que la nova frase soni com la primera. En un test d'avaluació subjectiva s'ha mostrat com les emocions són reconeixibles en les frases transformades, i que tant les característiques prosòdiques com espectrals del senyal són importants a l'hora de proporcionar una determinada emoció a la veu.

Igualment estem treballant amb el sistema de conversió text-parla construint bases d'unitats diferenciades per a cada emoció i locutor. Això permet disposar d'una sèrie de veus en el sistema que es poden seleccionar amb comandes XML inserides en el text d'entrada per tal de forçar la base d'unitats amb la qual es sintetitza. Aquesta mateixa idea s'està portant a les funcions de selecció internes del sintetitzador, de manera que es puguin assignar un cost específic per a inclusió d'una determinada unitat en funció del tipus d'emoció de la frase de qual prové.

Tot sembla indicar que és possible aconseguir en un futur relativament curt de temps sistemes que incloguin la possibilitat de generar parla més espontània i amb variants emocionals. El desenvolupament de millors models de regles prosòdiques i de transformació de les característiques de la veu, com la creació de bases d'unitats específiques i tècniques de selecció, seran necessaris per a aconseguir millorar els sistemes de conversió text-parla.

Referències

Jordi Adell , Antonio Bonafonte (2004). "Towards phone segmentation for concatenative speech synthesis". 5th ISCA Speech Synthesis Workshop (Pittsburgh, USA)

Ignasi Esquerra, Antonio Bonafonte (2004), "Habla emocional mediante métodos de re-síntesis y selección de unidades", XIX Simposium de la URSI (Barcelona)

Helenca Duxans, Antonio Bonafonte, Alexander Kain, Jan van Santen (2004), "Including dynamic and phonetic information in voice conversion systems". Int. Conference on Spoken Language Processing, (Jeju Island, Korea)

Pablo D. Agüero, Antonio Bonafonte (2004), "Intonation Modeling for TTS using a Joint Extraction and Prediction Approach". 5th ISCA Speech Synthesis Workshop (Pittsburgh, USA)