D1.17. Analiza metodelor de control și adaptare automată a expresivității vorbitorilor în sistemele de sinteză text-vorbire

Aceste rezultate au fost obținute prin finanțare in cadrul Programului PN-III Proiecte complexe realizate în consorții CDI, derulat cu sprijinul MEN – UEFISCDI,
Cod: PN-III-P1-1.2-PCCDI-2017-0818, Contract Nr. 73 PCCDI/2018:
“SINTERO: Tehnologii de realizare a interfețelor om-mașină pentru sinteza text-vorbire cu expresivitate”

Denumirea organizației participante in proiect	Acronim organizație	Tip organizație	Rolul organizației în proiect (Coordonator/partener)
Institutul de Cercetări Pentru Inteligență Artificială “Mihai Drăgănescu”	ICIA	UNI	CO
Universitatea Tehnică din Cluj-Napoca	UTCN	UNI	P1
Universitatea Politehnica din București	UPB	UNI	P2
Universitatea "Alexandru Ioan Cuza" din Iași	UAIC	UNI	P3

MINISTERUL EDUCAȚIEI NAȚIONALE

UEFISCDI

Date de identificare proiect

Număr contract:	PN-III-P1-1.2-PCCDI-2017-0818, Nr. 73 PCCDI/2018
Acronim / titlu:	**„SINTERO: Tehnologii de realizare a interfețelor om-mașină pentru sinteza text-vorbire cu expresivitate”**
Titlu livrabil:	**D1.17 Analiza metodelor de control și adaptare automată a expresivității vorbitorilor în sistemele de sinteză text-vorbire**
Termen:	**Mai 2018**
Editor:	**Adriana Stan (Universitatea Tehnică din Cluj-Napoca)**
Adresa de eMail editor:	**Adriana.Stan@com.utcluj.ro**
Autori, in ordine alfabetică:	**Mircea Giurgiu, Adriana Stan**
Ofițer de proiect:	**Cristian STROE**

Rezumat

Acest document prezintă o scurtă analiză a metodelor de adaptare și control automat a expresivității în cadrul sistemelor de sinteză text-vorbire. Vor fi prezentate metodele principale disponibile în cadrul sistemelor de sinteză concatenative, a celor statistice bazate pe modele Markov și a sistemelor bazate pe rețele neuronale multistrat.

Cuprins

1. Introducere

Sistemele de sinteză text-vorbire au atins deja un nivel al calității vocii apropiat de cel al vocii naturale. Rămâne însă problema variabilității și a expresivității acestor voci. Această problemă este de actualitate și datorită faptului că expresivitatea sau prozodia nu poate fi evaluată în mod obiectiv printr-un set limitat de parametri și de cele mai multe ori depinde de starea emoțională a persoanei care o evaluează, precum și de fondul cultural, etnic, educațional șamd.

Diferitele tipuri de sisteme de sinteză: concatenative, parametric-statistice sau cele ce modelează direct forma de undă, permit un control al expresivității și prozodiei specific, în funcție de arhitectura sistemului. Secțiunile următoare vor prezenta succint metodele principale de control a expresivității și prozodiei acestor sisteme.

2. Controlul expresivității în cadrul sistemelor concatenative

Sistemele de sinteză concatenative au fost până de curând cele mai utilizate sisteme de sinteză pentru aplicațiile comerciale. Principiul de bază al acestora îl reprezintă înregistrarea unui corpus de voce extins, de înaltă calitate, iar apoi concatenarea unor segmente audio pentru a crea informația textuală furnizată la intrare. Aceste segmente au lunigimi variabile și pornesc de la nivel de fonem sau silabă și pot ajunge până la nivel de sintagmă.

Problema acestor tipuri de sisteme este faptul că informația audio nu este parametrizată sub nicio formă, astfel că pentru controlul expresivității este necesară manipularea formei de undă. La modul cel mai simplu, controlul expresivității putea fi făcut prin selectarea segmentelor audio de concatenat pe baza unei traiectorii prozodice predefinite sau estimate din text [Hunt96]. Această metodă însă se limita la utilizarea unor evenimente prozodice existente în corpusul de voce, fără a putea extinde aceste evenimente în funcție de variabilitatea textului de intrare.

Pentru a permite o mai mare variabilitate a expresivității în sistemele concatenative, Cea mai utilizată metodă pentru acest scop este PSOLA [Valbret92]. În cadrul acestei metode, cadre de analiză sincrone cu frecvența fundamentală sunt extrase și prelucrate pentru a obține o altă frecvență fundamentală sau durată a unui fonem sau a unei silabe. De cele mai multe ori această manipulare a formei de undă duce la artefacte -- segmente nenaturale de voce ce se adaugă problemelor apărute la concatenarea diferitelor segmente acustice extrase din contexte diferite.

3. Controlul expresivității în cadrul sistemelor probabilistice bazate pe modele Markov

Până în 2015, direcțiile de cercetare în cadrul sistemelor de sinteză text-vorbire se bazau în mod preponderent pe modele statistice, dintre care cele mai utilizate erau modelele Markov cu stări ascunse (Hidden Markov Model-based Speech Synthesis System - HTS) [Tokuda13]. Aceste sisteme modelează vorbirea la nivel de fonem, folosind diferite tipuri de parametrizare a formei de undă. Cea mai utilizată metodă de parametrizare sau vocoder este cel STRAIGHT [??] ce folosțe 4 seturi de parametri: coeficienți Mel-cepstrali, coeficienți de aperiodicitate, frecvența fundamentală (F0) și durata. Pentru controlul expresivității se pot astfel manipula în mod independent modelele pentru F0 și durata. Cu toate acestea, în cadrul sistemelor de tip HTS, problema expresivității a reprezentat o provocare suplimentară deoarece naturalețea vocii sintetizate este condiționată de utilizarea unor înregistrări audio ce conțineau o prozodie cât mai liniară, pentru ca modelele statistice să poată utiliza un număr cât mai mare de exemple fonetice pentru același context.

Există, însă, un număr mare de sistemele bazate pe modele Markov ce înglobează într-o formă sau alta un modul de control al prozodiei, iar cele mai importante dintre acestea vor fi enumerate în paragrafele următoare.

Adnotarea prozodică a informației textuale

Încă din primele versiuni ale sistemelor, în cadrul etichetelor contextuale utilizate în antrenarea modelelor acustice bazate pe modele Markov, s-a utilizat setul de adnotări prozodice ToBI [Zen09]. Cu ajutorul acestora, se puteau marca atât în setul de date de antrenare, cât și la evaluare evenimentele prozodice existente la nivelul fonemelor individuale. Însă, datorită necesității generării unei traiectorii continue pentru fluxul F0, vocea generată strict pe baza acestei metode era inexpresivă și monotonă [??]. O primă încercare de a permite o mai mare variabilitate a traiectoriei F0 și a duratei a fost cea prin care s-a utilizat așa numita variantă globală (en. global variance) [Toda05], însă rezultatele nu au fost cu mult mai expresive. Iar problema principală a reprezentat-o faptul că sistemele HTS nu foloseau informații suprasegmentale (de ex. la nivel de silabă sau de cuvânt). Problema majoră a acestei metode este faptul că la evaluarea sistemului (sinteza propriu-zisă), aceste adnotări prozodice trebuiau deduse direct din text, fapt ce era greu de realizat.

Controlul modelelor acustice

Pe lângă setul de etichete de bază utilizate pentru generarea semnalului vocal, pentru a controla mai ușor expresivitatea vocilor sintetizate, se pot adăuga anumite caracteristici la nivel supra segmental sau metalingvistic sau care să înglobeze elemente de realizare articulatorie. De exemplu, [Miyanaga04] și [Nose07] utilizează modele Markov ascunse cu regresie multiplă (en. Multiple Regression HMMs), în cadrul cărora mediile modelelor probabilistice sunt controlate la momentul sintezei prin intermediul unor caracteristici auxiliare, cum ar fi stilul de vorbire, emoția, etc. Pe de altă parte, [Ling09] folosește ca și parametri de control, adnotări ale mișcărilor organelor fonatoare. În cadrul acestei metode și având la dispoziție un modul ce permite extragerea parametrilor articulatori din semnalul audio, nivelul și tipul expresivității vocii sintetizate poate fi mult mai bine controlat.

Adaptarea modelelor acustice

În cazul în care există un set mai larg de voci înregistrate pentru un sistem de sinteză text-vorbire, există posibilitatea ca timbrul sau identitatea unui anumit vorbitor să poată fi combinată cu modelele prozodice sau intonația unui alt vorbitor. În acest caz, dacă înregistrările conțin aceeași informație textuală, o simplă înlocuire a arborilor de decizie aferenți frecvenței fundamentale și duratei poate fi suficient [SWARA17].

În cazul în care datele vorbitorului de la care se dorește transferul expresivității nu sunt suficiente, se poate realiza adaptarea modelelor acustice folosind diverse metode probabilistice, similare cu cele utilizate pentru adaptarea identității vorbitorului [Trueba15] sau prin modificarea modelelor Markov cu stări ascunse, cum ar fi factorizarea lor [Latorre14].

4. Controlul expresivității în sisteme de sinteză bazate pe rețele neuronale multistrat

Rețelele neuronale s-au impus ca și algoritm de învațare automată aproape universal începând cu anul 2006, când atât performanțele mașinilor de calcul au crescut precum și odată cu prezentarea algoritmului de învățare rapidă a ponderilor rețelelor de către G. Hinton [Hinton06]. Ca urmare, majoritatea problemelor de învățare neliniară și pentru care există suficiente date au fost rezolvate și îmbunătățite cu ajutorul rețelelor neuronale multistrat (en. Deep Neural Networks - DNN). Printre acestea, se numără și sistemele de recunoaștere [Hinton12] și sinteză a vorbirii [Zen13]. În partea de sinteză a vorbirii, sistemele bazate pe rețele neuronale au depășit cu mult naturalețea celor bazate pe modele probabilistice [Oord16]. Au

rămas încă deschise problemele de expresivitate a vocii sintetizate, precum și cea de creare a vocilor din seturi de date reduse, prin așa numita adaptare a vorbitorilor.

În ceea ce privește expresivitatea sistemelor de sinteză bazate pe rețele neuronale, abordările includ de cele mai multe ori extinderea setului de caracteristici de intrare, cu un set de caracteristici de prozodie sau de stil de vorbire. Problema majoră a acestor sisteme este necesitatea existenței unui corpus de voce de dimensiuni mari pe baza căruia să se realizeze antrenarea rețelei. În cazul în care acest corpus de voce nu este disponibil se poate utiliza învățarea prin transfer. Această metodă presupune pre-antrenarea rețelei cu un set de date amplu, care însă nu este proiectat specific pentru scopul dat, iar apoi rafinarea acestei rețele cu un set de date specific, dar de dimensiuni reduse [Sawada17].

În paragrafele următoare enumerăm principalele metode de realizarea a vocilor sintetizate expresive în cadrul celor mai cunoscute sisteme de sinteză text-vorbire bazate pe rețele neuronale multistrat.

Sistemul de sinteză bazat pe rețele neuronale Tacotron [Wang17] de la Google prezintă o extindere a funcțiilor sale de bază pentru a crea voci expresive, prin utilizarea unor caracteristici latente învățate din corpusul de antrenare și utilizate ulterior și la sinteză [Sherry-Ryan18]. Rezultatele lor prezintă și utilizarea unor seturi de parametri de control ai prozodiei dinafara setului de antrenare. Tot în cadrul Tacotron, [Stanton18] a incorporat o reprezentare latentă denumită Global Style Tokens și care poate fi generată automat din textul de intrare.

Cei de la Baidu au introdus un sistem denumit EMPHASIS [Li18] ce modelează dependențele dintre caracteristicile lingvistice și cele acustice folosind o rețea de regresie. Caracteristicile acustice sunt, de asemenea, grupate astfel încât să se maximizeze caracteristicile de emotivitate și prozodie.

Rezultatele ambelor sisteme de sinteză sunt de o calitate foarte bună, rămâne însă deschisă problema generării automate a etichetelor de expresivitate și prozodie automat din text, precum și transferul stilului de vorbire din date cât mai puține.

5. Concluzii

Acest document a prezentat succint cele mai importante metode de analiză și control a expresivității în cadrul sistemelor de sinteză text-vorbire. Aceste metode au fost clasificate în funcție de tipul sistemului în cadrul căruia au fost aplicate, în sisteme statistice bazate pe modele Markov și sisteme bazate pe rețele neuronale multistrat.

Este evident faptul că studiul expresivității vocii umane este în continuare un subiect de cercetare important, dat fiind și faptul că evaluarea expresivității este mai degrabă o evaluare subiectivă, fără a fi dependentă în mod clar de anumiți parametri măsurabili. Există, însă, un oarecare nivel de acord între evaluatori în ceea ce privește, de exemplu, realizarea vocală a unor emoții puternice, cum ar fi mânia sau bucuria.

Totodată, nivelul de expresivitate sau prozodia unui vorbitor particular poate fi transpusă unei voci. Din nou, studiile din acest domeniu nu pot să identifice clar un set de parametri sau măsuri obiective a ceea ce reprezintă un anumit stil oratoric. Se pot identifica, însă, un număr redus de modificări de durată sau variații relative ale F0.

Se mai pune problema și de emfază a unor cuvinte ce trebuie accentuate pentru a transmite mesajul în mod clar către ascultător. Această emfază depinde de obicei de vorbitor, precum și de contextul mai larg al discursului din care face parte propoziția sau fraza curentă. Aceste evaluări la nivel de dialog sau discurs amplu țin mai degrabă de analiza textului și identificarea acestei emfaze.

6. Bibliografie

Hinton06 G. Hinton, S Osindero, YW Teh, *A fast learning algorithm for deep belief nets,* Neural Computation, 2006

Hinton12 Geoffrey Hinton, Li Deng, Dong Yu, George Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara Sainath, and Brian Kingsbury, *Deep Neural Networks for Acoustic Modeling in Speech Recognition*, IEEE Signal Processing Magazine, vol 29, issue 6, pp 82-97

Hunt96 AJ Hunt, AW Black, *Unit selection in a concatenative speech synthesis system using a large speech database,* Proc. Of ICASSP, 1996

Latorre14 J Latorre, V Wan, J Yanagisawa, *Voice expression conversion with factorised HMM-TTS models,* Proc. Of Interespeech 2014

Li18 H Li, Y Kang, Z Wang, *EMPHASIS: An emotional phoneme-based acoustic model for speech synthesis system,* Proc of Interspeech 2018

Ling09 ZH Ling, K Richmond, J Yamagishi, RH Wang, *Integrating articulatory features intro HMM-based parametric speech synthesis,* IEEE Trans Audio Speech Language, vol 17, no 6, pp 1171-1185

Miyanaga04 K Miyanaga, T Masuo, T Kobayashi, *A style control technique for HMM-based speech synthesis*, Proc of Interspeech 2004

Nose07 T Nose, J Yamagishi, T Masuko, T Kobayashi, *A style control technique for HMM-based expressive speech synthesis,* IEICE Trans Inf Syst, vol 90D, no 9, pp 1406-1413, 2007

Oord16 Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, Koray Kavukcuoglu, *WaveNet: A Generative Model for Raw Audio,* arXiv 1609.03499

Sawada17 Yoshihide Sawada, Yoshikuni Sato, Toru Nakada, Kei Ujimoto, Nobuhiro Hayashi, *All-Transfer Learning for Deep Neural Networks and its Application to Sepsis Classification,* arXiv 1711.04450

Skerry-Ryan18 RJ Skerry-Ryan, Eric Battenberg, Ying Xiao, Yuxuan Wang, Daisy Stanton, Joel Shor, Ron J. Weiss, Rob Clark, Rif A. Saurous, *Towards End-to-End Prosody Transfer for Expressive Speech Synthesis with Tacotron,* arXiv 1803.09047

Stanton18 D Stanton, Y Wang, RJ Skerry-Ryan, *Predicting expressive speaking style from text in end-to-end speech synthesis,* arXiv 1808.01410

SWARA17 A. Stan, M. Giurgiu, Demonstrator online pentru combinarea parametrilor modelelor HTS: http://romaniantts.com/swmix/

Toda05 T Toda, K Tokuda, *Speech Parameter Generation Algorithm Considering Global Variance for HMM-based Speech Synthesis,* Proc. Interspeech 2015

Trueba15 J Lorenzo-Trueba, R Barra-Chicote, R San-Segundo, J Ferreiros, J Yamagishi, JM Montero, Emotion transplantaion through adaptation in HMM-based speech

synthesisi, Computer Speech and Language, vol 34, pp 292-307, 20015

Valbret92 H.Valbret, E.Moulines, J.P.Tubach, Voice transformation using PSOLA technique, Speech Communication, vol 11, issues 2-3, pp 175-187, 1992

Veaux11 C. Veaux, X Rodet, *Prosodic control of unit-selection speech synthesis: A probabilistic approach,* Proc. Of ICASSP 2011.

Wang17 Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, Rif A. Saurous, *Tacotron: Towards End-to-End Speech Synthesi*s, arXiv 1703.10135

Zen09 H Zen, K Tokuda, A Black, *Statistical parametric speech synthesis,* Speech Communication, vol 51, no 11, pp 1039-1064, 2009

Zen13 H Zen, A Senior, M Schuster, *Statistical Parametric Speech Synthesis Using Deep Neural Networks*, Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), IEEE (2013), pp. 7962-7966