Universitatea Babeș-Bolyai
Facultatea de Matematică și Informatică

CARACTERISTICI LEXICALE AUGMENTATE ȘI APLICAȚII DE ADAPTARE CU CANTITĂȚI LIMITATE DE DATE PENTRU SINTEZĂ TEXT VORBIRE ÎN LIMBA ROMÂNĂ

Rezumatul tezei de doctorat

Student doctorand:
Beáta LÁZÁR-LŐRINCZ

Conducător științific:
Prof. dr. Bazil PÂRV

2022

Cuprins

Cuprinsul tezei de doctorat

Lista publicațiilor

Conference proceedings

1. Beáta Lőrincz, Maria Nuțu, Adriana Stan, Romanian Part of Speech Tagging using LSTM Networks, In 2019 IEEE 15th International Conference on Intelligent Computer Communication and Processing (ICCP), IEEE, pp. 223-228, 2019.

   Rang C, 2 puncte.

2. Maria Nuțu, Beáta Lőrincz, Adriana Stan, Deep learning for automatic diacritics restoration in Romanian, In 2019 IEEE 15th International Conference on Intelligent Computer Communication and Processing (ICCP), IEEE, pp. 235-240, 2019.

   Rang C, 2 puncte.

3. Beáta Lőrincz, Maria Nutu, Adriana Stan, Mircea Giurgiu, An Evaluation of Postfiltering for Deep Learning Based Speech Synthesis with Limited Data, In 2020 IEEE 10th International Conference on Intelligent Systems (IS), IEEE, pp. 437-442, 2020.

   Rang C, 1 punct.

4. Beáta Lőrincz, Concurrent phonetic transcription, lexical stress assignment and syllabification with deep neural networks, In Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 24th International Conference KES2020, Procedia Computer Science, 176, pp. 108-117, 2020.

   Rang B, 4 puncte.

5. Beáta Lőrincz, Adriana Stan, Mircea Giurgiu, Speaker verification-derived loss and data augmentation for DNN-based multispeaker speech synthesis, In 2021 29th European Signal Processing Conference (EUSIPCO), IEEE, pp. 26-30, 2021.

   Rang B, 4 puncte.

6. Beáta Lőrincz, Adriana Stan, Mircea Giurgiu, An objective evaluation of the effects of recording conditions and speaker characteristics in multi-speaker deep neural speech synthesis, In Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 25th International Conference KES2021, Procedia Computer Science, 192, pp. 756-765, 2021.

   Rang B, 4 puncte.

7. Stefan Daniel Dumitrescu, Petru Rebeja, Beáta Lőrincz, Mihaela Gaman, Andrei-Marius Avram, Mihai Ilie, Andrei Pruteanu, Adriana Stan, Lorena Rosia, Cristina Iacobescu, Luciana Morogan, George-Andrei Dima, Gabriel Marchidan, Traian Rebedea, Madalina Chitez, Dani Yogatama, Sebastian Ruder, Radu Tudor Ionescu, Razvan Pascanu, Viorica Patraucean LiRo: Benchmark and leaderboard for Romanian language tasks, In

Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1 pre-proceedings (NeurIPS Datasets and Benchmarks 2021), 2021.

Rang A*, 0.6 puncte.

8. Beáta Lőrincz, Bogdan Iudean, Andreea Vescan, Experience report on teaching testing through gamification, In Proceedings of the 3rd International Workshop on Education through Advanced Software Engineering and Artificial Intelligence, pp. 15-22, 2021.

   Rang A, 8 puncte.

9. Beáta Lőrincz, Contributions to neural speech synthesis using limited data enhanced with lexical features, In Proceedings of the 1st ISCA Symposium on Security and Privacy in Speech Communication, pp. 83-85, 2021.

   Rang D, 1 punct.

10. Adriana Stan, Beáta Lőrincz, Maria Nutu, Mircea Giurgiu, The MARA corpus: Expressivity in end-to-end TTS systems using synthesised speech data, In 2021 International Conference on Speech Technology and Human-Computer Dialogue (SpeD), IEEE, pp. 85-90, 2021.

    Rang D, 0 punct.

Book chapters

1. Adriana Stan, Beáta Lőrincz, Generating the Voice of the Interactive Virtual Assistant, In Virtual Assistant, IntechOpen, 2021.

   Rang D, 1 punct.

Scorul publicațiilor: 27.6 puncte.

1. Introducere

Sistemele de sinteză text-vorbire sunt utilizate în mod obişnuit în viaţa de zi cu zi, încorporate în aplicaţii de asistenţi virtuali, sisteme de navigaţie sau orice aplicaţie software, care permite ca conţinutul său textual să fie convertit în vorbire sintetizată. Odată cu răspândirea tehnologiilor de învăţare profundă, soluţiile bazate pe reţele neuronale sunt frecvent angajate pentru astfel de aplicaţii, deoarece vorbirea sintetizată atinge o naturaleţe apropiată de cea a vorbirii umane.

Însă există încă o serie de probleme care prezintă provocări pentru sinteza vorbirii bazată pe reţele neuronale. Pentru a obţine vorbire sintetizată cu o naturaleţe aproape de vorbirea umană, sunt necesare cantităţi mari de date audio de înaltă calitate pentru antrenament. Acest lucru este valabil atât pentru sistemele de sinteză care includ un singur vorbitor cât şi pentru identităţi multiple de vorbitori. Naturaleţea este importantă, deoarece sistemele de sinteză care generează vorbire inteligibilă, dar au un sunet mai puţin natural, nu sunt fezabile pentru aplicaţiile care generează segmente de vorbire mai lungi. Obţinerea unei similitudini ridicate a vorbitorului, între ieşirea sintetizată şi vocea umană înregistrată pentru datele de antrenament, este deosebit de benefică pentru sistemele de sinteză care urmăresc să recreeze vocile pentru cei care şi-au pierdut capacitatea de a vorbi.

Pe lângă naturaleţea şi similitudinea vorbitorului, expresivitatea vorbirii sintetizate contribuie la eficienţa şi aplicabilitatea acestuia, deoarece transmite informaţii suplimentare despre intonaţia şi expresiile asociate conţinutului textual. Sinteza expresivă a vorbirii este potrivită în special pentru naraţiunea cărţilor audio. Modelarea caracteristicilor prozodice, care rezultă într-o ieşire de vorbire sintetizată expresivă, reprezintă în prezent o provocare pentru sistemele de sinteză a vorbirii bazată pe reţele neuronale.

Comunităţile de cercetare şi industriale au publicat un număr mare de arhitecturi neuronale, care propun modele pentru sinteza vorbirii. Majoritatea acestor modele se concentrează pe limbile mainstream, cum ar fi engleza, dar limbile cu resurse reduse, precum limba română, încă suferă din cauza lipsei de resurse şi instrumente care sunt capabile să genereze vorbire sintetizată de înaltă calitate.

Întrucât una dintre principalele probleme de mare interes ale sintezei de vorbire bazată pe reţele neuronale este necesitatea unor cantităţi mari de date de înaltă calitate pentru antrenarea sistemelor, în această lucrare este abordată problema cantităţilor de date reduse folosite pentru antrenare. Acest scenariu este examinat, atât într-o configuraţie de sinteză a vorbirii cu un singur vorbitor, cât şi cu mai mulţi vorbitori. Metodele prezentate urmăresc să obţină naturaleţea şi similitudinea ridicată a vorbitorilor pentru modele antrenate pe limba română.

Este analizat controlul asupra caracteristicilor prozodice ale vorbirii de ieşire, pentru a îmbunătăţi efectul expresiv al vorbirii sintetizate. Din acest motiv, este introdus un nou set de date expresive pentru limba română. Caracteristicile lingvistice sunt examinate în acest

proces, textul de intrare este îmbunătăţit cu caracteristici lingvistice pentru a examina efectul acestor informaţii suplimentare asupra vorbirii sintetizate.

Obiectivul principal al tezei cuprinde îmbunătăţirea naturaleţei şi similitudinii vorbitorului pentru vorbirea sintetizată, folosind caracteristici lexicale augmentate pentru textul de intrare şi cantităţi reduse de date audio pentru sistemele de sinteză a vorbirii bazată pe reţele neuronale, axate pe limba română.

1.1. Structura tezei

Teza este structurat după cum urmează.

Capitolul 1 descrie principalele motivaţii şi obiective ale tezei**.**

Capitolul 2 prezintă concepte legate de generarea vorbirii şi analizează cele două componente majore ale sintezei vorbirii: procesarea textului şi modelarea acustică. Detaliem componentele şi conceptele aferente, contextul istoric şi metodele actuale de ultimă generaţie pentru sinteza vorbirii. Capitolul se încheie cu enumerarea metodelor de evaluare a modelelor de sinteză a vorbirii.

Capitolul 3 descrie instrumentele de procesare a textului care contribuie la componentele front-end ale sistemelor de sinteză text-vorbire. Sunt discutate părţile de vorbire, transcrierea fonemică, silabificarea şi accentul lexical, care pot fi utilizate pentru a îmbunătăţi textul de intrare al sistemelor de sinteză a vorbirii. Adiţional, menţionăm o analiză a sentimentelor, un experiment de restaurare a diacriticelor şi contribuţii la un sistem de referinţă pentru sarcinile de evaluare a limbii române.

Capitolul 4 detaliază principalele contribuţii ale acestei teze, deoarece prezintă experimentele efectuate cu diferite arhitecturi de sinteză a vorbirii. Prima secţiune a capitolului prezintă metoda de postfitering utilizată pentru îmbunătăţirea rezultatului unei reţele neuronale feed-forward prin angajarea adaptării de vorbitor. Următoarea secţiune prezintă două experimente cu mai mulţi vorbitori evaluate pe arhitecturile actuale de ultimă generaţie şi propuneri de îmbunătăţire şi evaluare a rezultatelor acestor sisteme antrenate pe limba română. Acest capitol discută, de asemenea, introducerea unui corpus de vorbire expresivă pentru limba română şi experimente cu sisteme de sinteză menite să ofere un control interactiv asupra trăsăturilor prozodice ale vorbirii sintetizate. Tehnicile de adaptare a vorbitorilor sunt discutate în diferite secţiuni ale capitolului.

Teza se încheie cu **Capitolul 5** care rezumă principalele concluzii şi enumeră propunerile de dezvoltare viitoare.

1.2. Contribuții originale

Contribuţiile tezei sunt rezumate mai jos cu numărul de capitol corespunzător.

În Capitolul 2 a fost elaborat un set de termeni referitori la procesarea digitală a vorbirii. Au fost detaliate componenta de procesare a textului, care converteşte textul introdus în caracteristici textuale utilizate pentru generarea vorbirii, şi componenta de modelare acustică, inclusiv o prezentare istorică a sistemelor de sinteză a vorbirii. Această prezentare generală a oferit elementele fundamentale de procesare a vorbirii şi a fost folosit ca punct de plecare pentru experimentele noastre.

Îmbunătăţirea textului introdus cu caracteristici lexicale poate îmbunătăţi calitatea vorbirii sintetizate. Pe baza acestei observaţii, în Capitolul 3 am prezentat experimente care prezic partea de vorbire a secvenţei de intrare, transcrierea fonemică, silabificarea şi accentul lexical. Furnizarea acestor informaţii aplicaţiilor de procesare a vorbirii poate contribui la expresivitatea acesteia, iar mai ales în cazul limbilor fonetice, transcrierea fonemică poate contribui la o mai bună pronunţie a textului introdus. Experimentele suplimentare prezentate includ restaurarea diacriticelor care poate reintroduce simbolurile diacritice ale textului de intrare, analiza sentimentelor care poate fi folosită pentru a detecta starea afectivă a textului de intrare. Contribuţiile la sistemul de referinţă pentru sarcini lingvistice prezentate încurajează cercetarea limbajului natural pentru limba română.

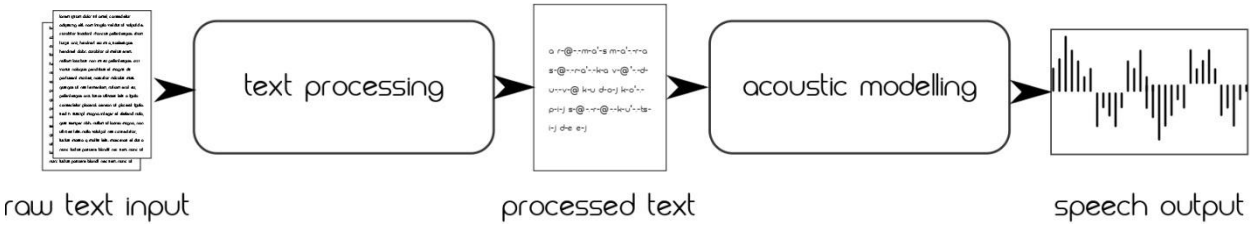
Sistemele de sinteză text-vorbire au atins o naturaleţe apropiată de cea a vorbirii umane, totuşi acestea se bazează pe cantităţi mari de date de vorbire de înaltă calitate. În capitolul 4 prezentăm experimente care se străduiesc să îmbunătăţească calitatea vorbirii de ieşire folosind o cantitate redusă de date. Au fost prezentate metodele de adaptare a vorbitorului, angajarea unei reţele de verificare a vorbitorilor, augmentarea datelor, selecţia atentă a datelor vorbitorilor care urmăresc să producă vorbire sintetizată de înaltă calitate pentru limba română. Pentru a contribui la naturaleţea vorbirii de ieşire am prezentat şi experimente care vizează controlul prozodic al vorbirii sintetizate.

Cuvinte cheie: sinteză text-vorbire, sinteză bazată pe reţele neuronale, procesare de text, reţele neuronale profunde, limba română

2. Fundamente teoretice

În acest capitol discutăm componentele şi metodele utilizate pentru sistemele de sinteză text-vorbire. Detaliem etapele procesului pornind de la procesarea textului până la modelarea acustică pentru a ajunge la vorbirea sintetizată. Metodele utilizate pentru acest flux sunt examinate pornind de la abordări bazate pe reguli până la modelele de ultimă generaţie bazate pe reţele neuronale profunde.

Sinteza text-vorbire se referă la procesul de generare a semnalului de vorbire din text. Pornind de la textul de intrare prin mai mulţi paşi intermediari, este generat semnalul acustic sau forma de undă de vorbire. O prezentare generală a sintezei text-vorbire este ilustrată în următoarea figură:



La prima vedere, acest proces pare o mapare directă de la caracterele textului introdus până la realizarea lor acustică. Cu toate acestea, există o serie de probleme tehnice pentru care această problemă devine extrem de complexă, cum ar fi variabilitatea vorbitorului şi a vorbirii sau modelarea caracteristicilor prozodice.

Înainte de a discuta contextul istoric al sintezei vorbirii, definim un set de termeni legaţi de generarea vorbirii, reprezentările acesteia şi provocările impuse de acest proces. Descrierea elementelor fundamentale ale vorbirii este apoi urmată de secţiuni care descriu cele două componente majore ale sistemelor de sinteză text-vorbire. Abordările utilizate pentru acestea în ultimele decenii sunt detaliate. Capitolul se încheie cu descrierea metodelor utilizate pentru evaluarea sistemelor de sinteză text-vorbire.

Acest capitol se bazează pe capitolul de carte publicat în [1].

3. Contribuții la procesarea textului

Prima componentă a sistemului de sinteză text-vorbire include etapele de procesare a textului, care transformă textul de intrare într-o reprezentare, care va fi folosită pentru modelarea acustică. Deoarece rezultatul sistemului de sinteză este realizarea acustică a textului de intrare, caracteristicile textuale afectează vorbirea sintetizată. Augmentarea textului introdus cu informații lexicale sau caracteristici legate de conținut poate îmbunătăți vorbirea sintetizată în ceea ce privește expresivitatea și naturalețea.

Complexitatea etapelor de procesare a textului este foarte dependentă de limbajul folosit; din acest motiv, la selectarea acestor pași, trebuie luate în considerare caracteristicile limbajului.

În acest capitol, prezentăm sarcini de procesare a textului care pot fi utilizate ca funcții de augmentare pentru textul de intrare a sistemelor sinteză text-vorbire. Prima sarcină descrisă urmărește să prezică partea de vorbire a cuvintelor. Adăugarea părților de vorbire pentru cuvintele textului introdus este utilă pentru dezambiguizarea cuvintelor care sunt scrise în același mod, dar care sunt pronunțate diferit. A doua secțiune prezintă soluții concurente pentru sarcinile de transcriere fonetică, atribuire a accentului lexical și silabificare. Transcrierea fonetică este avantajoasă pentru limbile bogate din punct de vedere fonetic, în timp ce accentul lexical și limitele silabelor poartă informații despre intonația și ritmul vorbirii. Sarcini suplimentare prezentate sunt restaurarea semnelor diacritice și analiza sentimentelor. Pentru limbile care conțin simboluri diacritice, precum limba română, este important să fie incluse în textul de intrare, pentru a păstra sensul original al textului. Analiza sentimentelor oferă informații despre starea afectivă a textului și poate fi folosită pentru a eticheta textul introdus cu sentimente.

Analizăm aceste sarcini datorită importanței lor în procesarea vorbirii. La finalul capitolului, este prezentat un sistem de referință pentru evaluarea sarcinilor de procesare a textului, axat pe limba română.

3.1. Determinarea părții de vorbire

Determinarea părții de vorbire este una dintre sarcinile cheie ale procesării limbajului natural. Se referă la identificarea părții de vorbire a unui cuvânt și, opțional, proprietăți gramaticale suplimentare inerente unei părți de vorbire. Împreună cu alte componente ale procesării limbajului natural, cum ar fi lematizarea, silabificarea, detectarea limitelor cuvintelor sau propoziției și multe altele, își propune să ajute mașinile să înțeleagă și să proceseze limbajul natural.

Dificultatea sarcinii constă în faptul că aceeași formă ortografică a unui cuvânt poate avea un sens diferit, în funcție de context (omografi). O altă problemă este că declinația și

inflexiunile cuvintelor nu sunt regulate, mai ales în limbile bogate din punct de vedere morfologic, cum ar fi cazul limbii române.

Ca urmare, pentru a defini partea de vorbire corect al unui cuvânt în afară de ortografie, trebuie să luăm în considerare şi legăturile semantice dintre cuvinte.

Pentru sarcina de predicţie a părţilor de vorbire, examinăm modele de reţele neuronale compuse din straturi recurente şi le comparăm cu modele secvenţă-la-secvenţă care conţin codificatoare şi decodoare recurente. Cea mai mare acurateţe obţinută pentru determinarea părţii de vorbire rădăcină este de 99,18% şi 98,25% pentru părţile de vorbire extinse care conţin informaţii gramaticale suplimentare lângă partea de vorbire rădăcină.

3.2. Transcrierea fonetică, silabificarea şi determinarea accentului lexical

Alte instrumente de procesare a textului utilizate frecvent pentru sinteză text-vorbire sunt sarcinile de transcriere fonetică, silabificare şi determinarea accentului lexical.

Transcrierea fonetică sau conversia grafem-la-fonem se referă la procesul de reprezentare a formei scrise a unui cuvânt într-o formă derivată acustic, potrivită pentru o citire sau pronunţie corectă a cuvântului respectiv. Determinarea accentului lexical se referă la identificarea celei mai proeminente silabe dintr-un cuvânt. Silabificarea este procesul de împărţire a unui cuvânt în unităţi mai mici, în general aparţinând unui singur bloc de foneme de pronunţie.

Aceste sarcini depind în mare măsură de limba utilizată şi sunt vitale în multe fluxuri de procesare a vorbirii şi a limbajului. Cunoaşterea modului în care sunt pronunţate cuvintele este esenţială pentru producerea de sinteză text-vorbire şi sisteme de recunoaştere a vorbirii de înaltă calitate [2]. Pronunţia exactă a unui cuvânt într-o limbă depinde în mare măsură de fonemele componente ale acestuia. Cu toate acestea, structura silabelor şi accentul lexical este important în atribuirea sensului corect şi a accentuării acesteia [3].

Aceste sarcini au fost analizate şi o soluţie concurentă a fost propusă pentru rezolvarea lor. Necesitatea soluţiei concurente provine dintr-un aspect important al sintezei text-vorbire, care este procesarea în timp real. Având module separate pentru fiecare dintre sarcini se traduce într-o suprasarcină pentru fluxul de procesare. Ca contramăsuri, soluţia concurentă este propusă pentru transcrierea fonetică, determinarea accentului lexical şi silabificarea. Scopul este de a evalua fezabilitatea unui singur model care prezice împreună toate cele trei informaţii lingvistice. Limbile folosite în evaluare sunt engleza şi româna.

În ceea ce priveşte abordarea algoritmică, studiile recente au arătat că modelele secvenţă-la-secvenţă bazate pe reţele neuronale profunde sunt extrem de eficiente în rezolvarea sarcinilor de transcriere fonetică, atribuire a accentului lexical şi silabificare în diferite limbi [4].

Pornind de la această observaţie, studiul actual foloseşte arhitecturi de reţele neuronale recurente, de convoluţie şi de atenţie ca strategii principale de antrenament şi evaluează mai multe combinaţii şi variaţii ale unui scenariu de învăţare secvenţă la secvenţă. Sunt prezentate două studii de caz, primul se bazează pe arhitecturi recurente şi convoluţionale combinate cu straturi de atenţie, în timp ce al doilea se bazează pe arhitectura Transformer. Cele mai bune rezultate pentru primul studiu de caz au fost obţinute cu o combinaţie de straturi de convoluţie şi atenţie, unde acurateţea predicţiei comune pentru cele trei sarcini a fost de 58,96% pentru engleză şi 86,64% pentru română. Al doilea set de experimente obţine rezultate îmbunătăţite, cu o rată de eroare la nivel de cuvânt de 5,6% pentru setul de date românesc (pe un set de date mare românesc recent introdus numit RoLEX) şi o rată de eroare la nivel de cuvânt de 24,11% pentru setul de date în limba engleză.

3.3. Analiza sentimentelor

Etichetarea textului de intrare cu sentimente poate fi folosită ca o caracteristică suplimentară pentru sistemele de sinteză text-vorbire pentru a condiţiona vorbirea sintetizată pe baza sentimentului textului de intrare. Analiza sentimentelor este o sarcină fundamentală a procesării limbajului natural care determină starea afectivă a textului, folosind cel mai frecvent pozitiv şi negativ ca etichete. Această sarcină a fost abordată cu diverse metode bazate pe reguli, statistice sau de învăţare automată.

Aplicaţiile analizei sentimentelor sunt diverse; pe lângă aplicaţia sa în sistemele de sinteză expresivă a vorbirii, poate fi folosit pentru monitorizarea reţelelor sociale sau orice alte sarcini de analiză a textului, unde sentimentul textului poartă o informaţie valoroasă pentru aplicaţie. În cercetarea noastră, analiza sentimentelor este aplicată pentru a evalua părerile studenţilor legat de activităţile educaţionale.

3.4. Restaurarea semnelor diacritice

Sarcina de restaurare automată a diacriticelor se referă la reintroducerea simbolurilor diacritice ale unui text care nu conţine aceste simboluri. Acest proces de restaurare este util şi necesar deoarece multe dintre conţinutul electronic disponibil online nu conţin simboluri diacritice.

În cazul mai multor limbi, utilizarea semnelor diacritice poate schimba sensul şi pronunţia cuvintelor. Din acest motiv, textul de intrare al sistemelor de sinteză text-vorbire trebuie să conţină simbolurile diacritice pentru a facilita învăţarea corectă a pronunţiilor cuvintelor în vorbirea sinetizată.

În experimentele efectuate au fost prezentate arhitecturi de reţele neuronale compuse din straturi recurente şi convoluţionale pentru sarcina de restaurare a diacriticelor. Cele mai bune rezultate de acurateţe obţinute la nivel de cuvânt au fost de 97%.

3.5 Bias de gen în reprezentarea vectorială a cuvintelor

Reprezentări vectoriale pentru cuvinte sunt frecvent utilizate în aplicațiile de procesare a limbajului natural. Cuvintele sunt reprezentate prin vectori cu valori reale, astfel încât cuvintele cu semnificație similară sunt mai apropiați în spațiul vectorial. Cu toate acestea, pentru limbile cu gen gramatical, încorporarea cuvântului poate codifica un bias de gen moștenită din corpusul de antrenare. Pentru a elimina consecințele nedorite, s-au depus eforturi pentru reducerea și analizarea bias-ului de gen existente în modelele lingvistice [5]. Substantivele din limba română poartă un gen gramatical fiind clasificate în substantive masculine, feminine sau neutre. Multe studii care se concentrează pe atenuarea bias-ului de gen vizează limba engleză, dar metodele propuse nu pot fi transferate cu ușurință în alte limbi cu gen gramatical. Urmăm propunerea lui [5] și reproducem rezultatele prezentate pentru spaniolă și franceză pe limba română.

3.6. Sisteme de referință pentru sarcinile de evaluare lingvistică

Sisteme de referință pentru înțelegerea limbajului natural sau alte domenii încurajează cercetarea domeniului, oferă resurse pentru antrenarea și analiza sarcinilor și promovează standardizarea sarcinilor selectate. Platforma LiRo a fost propusă pentru a facilita cercetarea procesării limbajului natural pentru limba română. 10 sarcini de procesare a textului sunt disponibile în sistemul de referință și au fost create 3 noi seturi de date pentru a extinde resursele disponibile pentru limba română.

Rezultatele și experimentele prezentate în acest capitol se bazează pe lucrările originale publicate în [6-10].

4. Contribuţii la sisteme de sinteză text-vorbire

Cercetările recente în sinteza vorbirii se concentrează pe naturaleţea şi expresivitatea vorbirii sintetizate, pentru a creşte aplicabilitatea sistemelor de sinteză text-vorbire. Este rezonabil să admitem că, cu cât sună mai natural o voce sintetică, cu atât mai uşor şi mai plăcut este de folosit în orice mediu.

În acest capitol sunt abordate câteva dintre principalele provocări ale sistemelor de sinteză text-vorbire. Prima secţiune este dedicată analizei metodelor de îmbunătăţire a vorbirii sintetizate atunci când sunt disponibile date de cantităţi reduse. Acest caz este întâlnit frecvent în practică, în special pentru limbile cu resurse insuficiente, cum ar fi limba română. În general, scopul nostru este de a sintetiza voci cu cât mai puţine date posibil. A doua secţiune prezintă două seturi de experimente pentru sinteza vorbirii cu mai mulţi vorbitori, care vizează îmbunătăţirea vorbirii sintetizate, concentrându-se totodată pe disponibilitatea şi caracteristicile textului de intrare şi a datelor audio. Textul de intrare este completat cu informaţii lexicale bazate pe paşii de procesare a textului prezentaţi în capitolul anterior. Pentru experimente sunt utilizate diverse cantităţi de date audio, iar caracteristicile înregistrărilor audio sunt analizate în ceea ce priveşte calitatea vorbirii rezultate. Sunt de asemenea explorate metode de adaptare a vorbitorului cu diferite cantităţi de date de antrenament. Ultima secţiune discută expresivitatea vorbirii care contribuie în mare măsură la naturaleţea vorbirii sintetizate. Sunt examinate caracteristici precum prozodia şi ritmul care fac ca vorbirea să sune mai mult sau mai puţin uman.

4.1. Sinteză text-vorbire cu cantităţi reduse de date

În sinteza vorbirii bazată pe reţele neuronale, obţinerea unei vorbiri sintetizate de înaltă calitate se bazează pe disponibilitatea unor cantităţi mari de date de vorbire de înaltă calitate. Metodele de învăţare automată în general, inclusiv reţelele neuronale, se bazează pe cantităţi mari de date de antrenament. Când aceste date nu sunt uşor disponibile, se caută alte metode pentru a îmbunătăţi calitatea vorbirii sintetizate.

În acest capitol abordăm problema datelor reduse prin utilizarea metodei de postfiltrare realizată prin adaptarea vorbitorului. Aplicăm adaptarea vorbitorului pentru a îmbunătăţi calitatea vorbirii sintetizate. Rezultatele arată că, chiar şi cu doar 10 minute de înregistrări de vorbire, calitatea ieşirii vocale sintetice este îmbunătăţită prin procesul de postfiltrare.

4.2. Sisteme de sinteză text-vorbire folosind identități vocale multiple

Sistemele de sinteză text-vorbire sunt capabile să sintetizeze vorbirea pentru unul sau mai mulți vorbitori, ale căror identități sunt învățate în timpul procesului de antrenare. Sistemele de sinteză text-vorbire cu mai mulți vorbitori au avantajul de a încorpora mai multe identități vocale într-o singură rețea și permit utilizatorului să selecteze oricare dintre aceste identități fără schimbări de arhitectură. Acest scenariu permite, de asemenea, utilizarea a mai puține mostre per vorbitor, cu toate acestea, în modelul acustic rezultat, nu toți vorbitorii prezintă aceeași calitate sintetică. În acest studiu propunem metode pentru îmbunătățirea învățării identității vorbitorului a sistemelor de sinteză text-vorbire cu mai mulți vorbitori.

Primele două direcții explorate sunt: (1) forțarea rețelei să învețe o reprezentare mai bună a identității vorbitorului prin adăugarea unui termen suplimentar de funcție de cost bazat pe metrici de verificare a vorbitorului; și (2) augmentarea datelor de intrare referitoare la fiecare vorbitor utilizând metode de manipulare a formei de undă. Ambele metode sunt eficiente, termenul de funcție de cost suplimentar ajută la învățarea similitudinii de vorbitor, în timp ce augmentarea datelor îmbunătățește inteligibilitatea sistemului de sinteză text-vorbire cu mai mulți vorbitori.

A treia abordare exploatează un corpus paralel românesc care conține vorbire înregistrată în diferite condiții și evaluează dacă anumite caracteristici ale vocii sau condiții de înregistrare influențează calitatea rezultată a identității vocii respective într-un scenariu de sinteză text-vorbire cu mai mulți vorbitori. De asemenea, analizează dacă diferite reprezentări ale textului de intrare pentru sistemul bazat pe rețele neuronale de sinteză îmbunătățește sau deteriorează vorbirea sintetizată a diferiților vorbitori. Rezultatele au arătat că condițiile de înregistrare influențează cel mai mult calitatea vorbirii sintetizate.

4.3. Sinteză expresivă de text-vorbire

Sinteza expresivă de text-vorbire este de interes pentru comunități de cercetare, deoarece naturalețea sistemelor de sinteză este comparabilă cu cea a vorbirii umane [11], iar controlând prozodia, expresivitatea unor astfel de sisteme permite o gamă mai variată de aplicații.

În cazul modelelor end-to-end bazate pe rețele neuronale, reprezentarea prozodiei de vorbire este de obicei învățată ca o reprezentare internă în cadrul modelului. Deoarece prozodia nu este învățată în mod explicit de model, expresivitatea vorbirii sintetizate depinde și de corpusuri de vorbire utilizate ca date de antrenare pentru modelele bazată pe rețele neuronale. Cu toate acestea, resursele corpusurilor de vorbire pentru limbile non-mainstream precum limba română sunt limitate, iar lipsa resurselor împiedică procesul de dezvoltare a sintezei expresive de text-vorbire. Din acest motiv, am introdus un nou corpus

de vorbire în limba română numit MARA, prima lansare de corpus de vorbire expresivă românească. Folosind noul corpus de vorbire prezentat, a fost examinată utilizarea vorbirii sintetizate ca date de antrenare.

Rezultatele şi experimentele prezentate în acest capitol se bazează pe lucrările originale publicate în [12-15].

5. Concluzii și direcții viitoare de cercetare

În această teză, ne-am concentrat pe sinteza bazată pe rețele neuronale de text-vorbire și pe aspectele sarcinilor de procesare a limbajului natural care pot fi folosite pentru a îmbunătăți sistemele de sinteză. Experimentele noastre s-au bazat în mare parte pe seturi de date românești, deoarece scopul nostru a fost să contribuim la instrumentele și sistemele antrenate pe limbi cu resurse insuficiente, cum ar fi limba română. Pentru a fi comparabil cu literatura științifică, un număr selectat de experimente au fost prezentate folosind, de asemenea, seturi de date în limba engleză.

Înainte de a prezenta experimentele legate de procesarea textului și a vorbirii, am oferit o privire de ansamblu în Capitolul 2 despre fundamentele procesării vorbirii, elementele de procesare a textului relevante pentru procesarea vorbirii, respectiv sistemele de sinteză text-vorbire dezvoltate începând cu sinteza bazată pe reguli dinainte de anii '90 până la modelele actuale bazate pe rețele neuronale de ultimă generație.

Pornind de la această prezentare generală, în Capitolul 3 am prezentat un set de experimente legate de procesarea textului care pot fi încorporate în aplicațiile de sinteză de vorbire pentru a îmbunătăți vorbirea sintetizată. Primul experiment a prezis părțile de vorbire a cuvintelor care poate fi folosită ca caracteristică de augmentare a textului de intrare. Următorul experiment a prezentat modele pentru predicția concurentă a transcrierii fonemice, a silabificării și a accentului lexical al textului de intrare. În mod similar cu părțile de vorbire, această informație poate fi folosită pentru a îmbogăți informațiile textuale adăugate sistemelor de sinteză a vorbirii. Este descris un set suplimentar de experimente, inclusiv restaurarea diacriticelor și contribuția autorului la un sistem de referință pentru sarcinile de evaluare a limbii române.

Principalele contribuții ale acestei teze au fost elaborate în Capitolul 4. Primul studiu de caz prezentat propune să îmbunătățească sinteza de vorbire a sistemelor bazate pe rețele neuronale prin angajarea unei rețele de postfiltrare. După cum s-a subliniat în această secțiune, scopul a fost antrenarea sistemelor de sinteză pe o cantitate redusă de date și obținerea calității rezonabile pentru vorbirea sintetizată. Rețeaua de postfiltrare se bazează pe tehnica de adaptare a vorbitorului. Următoarea secțiune a prezentat două studii de caz de sinteză a vorbirii cu mai mulți vorbitori. Prima s-a bazat pe o arhitectură convoluțională și a angajat o rețea de verificare a vorbitorilor pentru a obține termeni pentru funția de cost suplimentară, respectiv metode de augmentare a datelor pentru a crește volumul datelor de antrenare. Al doilea set de experimente s-a concentrat pe evaluări obiective ale vorbirii sintetizate pe baza condițiilor de înregistrare și a caracteristicilor vorbitorului care pot fi utilizate pentru a evalua calitatea datelor de antrenare. Experimentele suplimentare prezentate enumeră contribuția autorului la crearea primului set de date expresive românești numit Mara și experimente care vizează controlabilitatea expresivității vorbirii.

Sinteza text-vorbire bazată pe rețele neuronale nu este considerată o problemă rezolvată în totalitate, deoarece ridică câteva probleme deschise care sunt de interes pentru comunitatea de cercetar. Una dintre probleme este legată de prozodie, deoarece vorbirea sintetizată, în special în cazul paragrafelor lungi, este necesar pentru a modela prozodia adecvată care nu este realizată de majoritatea sistemelor de sinteză de ultimă generație. Pentru a imita vorbirea spontană, vorbirea sintetizată ar trebui să modeleze și procesul mental de dezvoltare și generare a mesajului care ar putea include repetări sau pauze. În ceea ce privește adaptarea vorbitorului, sinteza bazată pe rețele neuronale a vorbirii oferă soluții pentru adaptarea rapidă, dar aceste metode pot fi îmbunătățite prin reducerea cantității de date de adaptare sau a timpului de antrenare.

Ca direcții viitoare de cercetare, ne propunem să continuăm lucrările asupra sistemelor de sinteză care încorporează modelarea prozodiei, și să permită sinteza condiționată de aspecte legate de prozodie, folosind limba română. Contribuțiile la adaptarea rapidă a vorbitorului vor fi, de asemenea, în centrul cercetării noastre, vizând dezvoltarea sistemelor de sinteză text-vorbire bazată pe rețele neuronale să fie adaptate la noi identități vocale cu o cantitate redusă de date.

De asemenea, intenționăm să lansăm sisteme de sinteză pentru limba română ca API, inclusiv modele antrenate pe arhitecturi convoluționale, recurente și bazate pe atenție. Crearea de corpusuri de text și vorbire va rămâne în interesul nostru, în special pentru limba română, care să faciliteze și să încurajeze munca altor cercetători din domeniu.

Bibliografie

[1] Adriana Stan and Beáta Lőrincz. Generating the voice of the interactive virtual assistant. In Virtual Assistant. IntechOpen, 2021.

[2] Daan van Esch, Mason Chua, and Kanishka Rao. Predicting Pronunciations with Syllabification and Stress with Recurrent Neural Networks. In INTERSPEECH, pages 2841–2845, 2016.

[3] Paul Taylor. Text-to-Speech Synthesis. Cambridge University Press, 2009.

[4] Benjamin Milde, Christoph Schmidt, and Joachim Kohler. Multitask Sequence-to-Sequence Models for Grapheme-to-Phoneme Conversion. In INTERSPEECH, pages 2536–2540, 2017.

[5] Pei Zhou, Weijia Shi, Jieyu Zhao, Kuan-Hao Huang, Muhao Chen, Ryan Cotterell, and Kai-Wei Chang. Examining gender bias in languages with grammatical gender. arXiv preprint arXiv:1909.02224, 2019.

[6] Beáta Lőrincz, Maria Nuțu, and Adriana Stan. Romanian part of speech tagging using LSTM networks. In 2019 IEEE 15th International Conference on Intelligent Computer Communication and Processing (ICCP), pages 223–228, 2019.

[7] Beáta Lőrincz. Concurrent phonetic transcription, lexical stress assignment and syllabification with deep neural networks. Procedia Computer Science, 176:108–117, 2020. Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 24th International Conference KES2020.

[8] Beáta Lőrincz, Bogdan Iudean, and Andreea Vescan. Experience report on teaching testing through gamification. In Proceedings of the 3rd International Workshop on Education through Advanced Software Engineering and Artificial Intelligence, pages 15–22, 2021.

[9] Maria Nuțu, Beáta Lőrincz, and Adriana Stan. Deep learning for automatic diacritics restoration in romanian. In 2019 IEEE 15th International Conference on Intelligent Computer Communication and Processing (ICCP), pages 235–240. IEEE, 2019.

[10] Stefan Daniel Dumitrescu, Petru Rebeja, Beáta Lőrincz, Mihaela Gaman, Andrei-Marius Avram, Mihai Ilie, Andrei Pruteanu, Adriana Stan, Lorena Rosia, Cristina Iacobescu, Luciana Morogan, George-Andrei Dima, Gabriel Marchidan, Traian Rebedea, Madalina Chitez, Dani Yogatama, Sebastian Ruder, Radu Tudor Ionescu, Razvan Pascanu, Viorica Patraucean, LiRo: Benchmark and leaderboard for Romanian language tasks, In Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1 pre-proceedings (NeurIPS Datasets and Benchmarks 2021), 2021.

[11] Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, R. J. Skerry-Ryan, Rif A.

Saurous, Yannis Agiomyrgiannakis, and Yonghui Wu. Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions. CoRR, abs/1712.05884, 2017.

[12] Beáta Lőrincz, Maria Nuțu, Adriana Stan, and Mircea Giurgiu. An evaluation of postfiltering for deep learning based speech synthesis with limited data. In 2020 IEEE 10th International Conference on Intelligent Systems (IS), pages 437–442. IEEE, 2020.

[13] Beáta Lőrincz, Adriana Stan, and Mircea Giurgiu. Speaker verification-derived lossand data augmentation for DNN-based multispeaker speech synthesis. In 2021 29th European Signal Processing Conference (EUSIPCO), pages 26–30, 2021.

[14] Beáta Lőrincz, Adriana Stan, and Mircea Giurgiu. An objective evaluation of the effects of recording conditions and speaker characteristics in multi-speaker deep neural speech synthesis. Procedia Computer Science, 192:756–765, 2021. Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 25th International Conference KES2021.

[15] Adriana Stan, Beáta Lőrincz, Maria Nuțu, and Mircea Giurgiu. The MARA corpus: Expressivity in end-to-end TTS systems using synthesised speech data. In 2021 International Conference on Speech Technology and Human-Computer Dialogue (SpeD), pages 85–90, 2021.