Studiu privind starea artei: Sisteme complementare de recunoaștere automată a vorbirii

Lucian Georgescu Horia Cucu Dan Oneață
Corneliu Burileanu Dragoș Burileanu

Cuprins

---

1 Introducere

Pentru dezvoltarea unui sistem de recunoaștere automată a vorbirii, resursa principală o reprezintă un corpus de vorbire pentru care există transcrierea corespunzătoare. Crearea unui model acustic puternic presupune antrenarea pe baza unui corpus de dimensiune cât mai mare, ce conține pronunții cât mai variate, provenite de la cât mai mulți vorbitori. Uneori, aceste resurse

audio adnotate lipsesc sau sunt dificil de procurat. Totuși, odată cu creșterea puterii computaționale și a capacității mediilor de stocare, interesul s-a mutat către corpusurile audio disponibile în online și mass-media. De exemplu, emisiuni de radio și televiziune, înregistrări ale ședințelor publice din cadrul anumitor instituții, lecturi, toate acestea sunt surse ușor accesibile și bogate în vorbire din viața reală, dar neadnotată.

Transcrierea unui astfel de corpus nu este deloc o sarcină trivială. Aceasta poate fi realizată în mod manual, prin ascultarea înregistrărilor, însoțită de scrierea simultană a textului aferent. Efortul în acest caz este considerabil, atât din punct de vedere al timpului necesar, cât și al costului. Mai mult, în cazul unor limbi slab dezvoltate, găsirea unor vorbitori nativi care să realizeze acest proces de adnotare manuală reprezintă în sine o problemă.

O altă modalitate de a obține un corpus de vorbire transcrisă, de data aceasta în mod automat, presupune chiar utilizarea unor sisteme de recunoaștere automată a vorbirii deja existente. Deoarece niciun astfel de sistem nu este perfect, iar transcrierile obținute conțin erori, un mod foarte întâlnit de tratare a acestei probleme presupune utilizarea concomitentă a mai multor sisteme. Figura 1 prezintă schema de principiu a unei astfel de abordări.

Transcrierile corpusului audio de la cele N sisteme trec printr-un proces de filtrare și selecție. Această etapă are sens numai în situația în care sistemele sunt diferite, iar erorile făcute de acestea sunt, de asemenea, distincte. Plecând de la această premisă, sunt considerate corecte părțile transcrise identic de către toate sistemele. Altfel, dacă cele N sisteme ar face erori simetrice, este echivalent cu utilizarea unui singur sistem, duplicat de N ori. În concluzie, o astfel de abordare este utilă numai dacă sunt folosite sisteme RAV complementare. Se poate spune despre două sau mai multe sisteme RAV că sunt complementare, dacă pentru aceeași secvență de vorbire, transcrierile obținute vor conține erori diferite.

În capitolele următoare vor fi prezentate principalele metode prin care se pot obține transcrieri complementare, împreună cu modul în care se realizează filtrarea acestora, în vederea adnotării cât mai precise a unor corpusuri de vorbire.

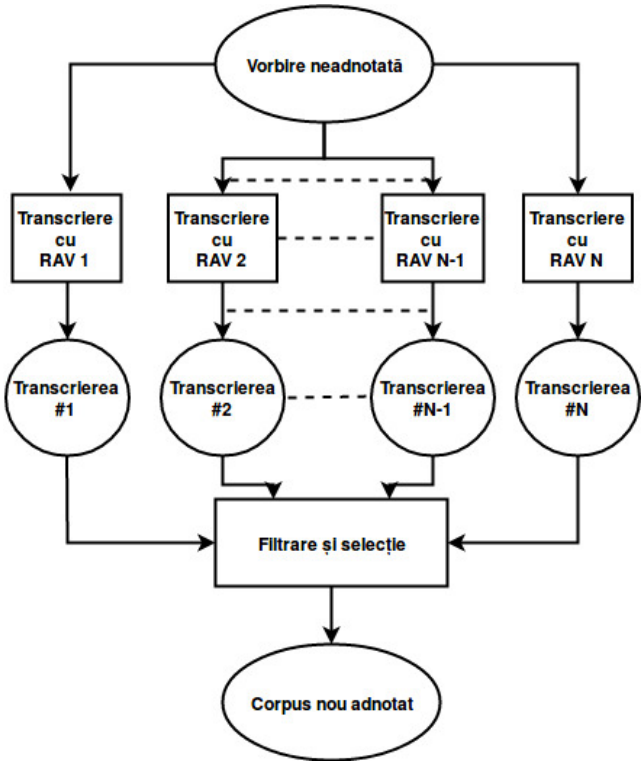

Figura 1: Adnotarea automată a corpusurilor audio

2 Metode de obținere a complementarității RAV

Obiectul acestui studiu este analiza complementarității sistemelor RAV, fiind dată de diferențe la nivelul:

- seturilor de date de antrenare;
- trăsăturilor acustice;
- algoritmilor de antrenare;
- algoritmilor de decodare;
- tipul modelului acustic.

Seturile de date de antrenare diferite (acustice/lingvistice) reprezintă un prim factor ce conduce la obținerea unor sisteme RAV complementare. Distincția dintre corpusuri poate fi dată de mai mulți factori:

- tipul vorbirii - citită sau spontană;
- mediul ambiant - în condiții de liniște sau zgomot;
- domeniul vorbirii - vorbire generală sau aparținând unei anumite sfere de activitate;
- modul de împărțire al corpusului în subseturi;
- genul vorbitorilor (masculin/feminin);
- limba utilizată;

Acești factori au un impact direct asupra diversității modelului acustic. Modelele acustice antrenate pe date care diferă în acest mod sunt complementare deoarece fiecare surprinde manifestări acustice diferite ale fonemelor. În (Koctúr et al., 2016) și (Koctúr et al., 2017) sunt utilizate două sisteme RAV complementare, antrenate pe seturi de date diferite. Scopul a fost transcrierea automată a unui corpus de vorbire achiziționat din transmisiuni de televiziune. În (Cucu et al., 2014) și (Cucu et al., 2015), ambele modele acustice sunt de același tip, dar ele au fost antrenate cu subseturi diferite din corpusul de antrenare: vorbire citită și vorbire spontană. Un alt element ce poate oferi transcrieri complementare în faza de decodare este modelul de limbă aplicat. În (Lojka and Juhár, 2014) se folosesc atât modele de limbă antrenate pe corpusuri generale de text, cât și modele de limbă antrenate pe text provenind dintr-un domeniu specific. Limba pentru care a fost antrenat modelul este de asemenea un element important, deoarece nu toate limbile folosesc același set de foneme, iar în cazul fonemelor ce se regăsesc în mai multe limbi, unele sunt perfect identice, în timp ce în cazul altora, se pot face asocieri de similitudine. Astfel de abordări, în care se folosesc modele acustice antrenate pe limbi diferite, se regăsesc în (Vu et al., 2010), (Vu et al., 2011), (Oparin et al., 2013) și (Haitao et al., 2016).

Trăsăturile extrase din semnalul vocal sunt un alt element ce duce la obținerea sistemelor RAV complementare. Așa cum este cunoscut, în sistemele RAV, nu se lucrează cu formă de undă brută, ci semnalul este parametrizat. Deși mai toate metodele de extracție a parametrilor vizează

păstrarea informației utile, cu rol discriminatoriu în producerea fonemelor, iar majoritatea operațiilor efectuate sunt comune, există mai multe tipuri de trăsături obținute din semnal. De exemplu, în (Lojka and Juhár, 2014) se folosesc atât coeficienți mel-cepstrali (MFCC), cât și coeficienți perceptuali de predicție liniară (PLP). Autorii utilizează în (Garau and Renals, 2008) atât coeficienți clasici, de tip MFCC și PLP, cât și coeficienți extrași prin metoda numită *pitch-synchronous analysis*, asupra cărora a fost aplicată normalizarea lungimii tractului vocal (VTLN). Mai mult, au fost încercate și diferite combinații între aceste trăsături. Cea mai simplă formă de combinare o reprezintă concatenarea vectorilor de trăsături acustice. Această operație se realizează prin aplicarea analizei discriminatorii liniare heteroscedastice (HLDA). Metoda este însă nefezabilă, deoarece dimensiunea spațiului ce trebuie modelat este crescută. Reducerea dimensionalității se realizează prin aplicarea unor tehnici ca analiza componentelor principale (PCA) sau analiza discriminatorie liniară (LDA). (Li et al., 2016) folosește două sisteme RAV complementare, ce diferă atât prin tipul modelului acustic, cât și prin tipul trăsăturilor vocale. Unul dintre sisteme este bazat pe coeficienți PLP, 13-dimensionali, asupra cărora se efectuează o operație de normare a mediei și variației (CMVN), în timp ce al doilea este bazat pe coeficienți mel 40-dimensionali, extrași cu bancuri de filtre. Aceleași tipuri de trăsături, împreună cu MFCC și combinații între acestea, sunt utilizate în (Shen et al., 2016).

Tipurile diferite de modele utilizate în sistemele RAV pot reprezenta surse de complementaritate. Astfel, în (Ma et al., 2012) sunt utilizate modele acustice de tip mixturi de gaussiene (GMM) sau subspații de mixturi gaussiene (SGMM), antrenate în mod convențional, plecând de la vectori de trăsături acustice, sau prin inițializare cu parametrii obținuți dintr-un model preantrenat de vorbire generală (UBM). Complementaritatea este dată în (Li et al., 2016) de utilizarea modelelor acustice de tip GMM, respectiv modele acustice antrenate cu rețele neuronale profunde (DNN). Aceleași tipuri de modele acustice sunt folosite și în (Huang et al., 2013). Autorii prezintă în (Shen et al., 2016) folosirea mai multor tipuri de modele acustice: SGMM, DNN sau rețele neuronale convoluționale (CNN-DNN). Tot rețele neuronale în diferite configurații sunt utilizate și în (Bell et al., 2013).

Algoritmii de antrenare ai modelelor fac ca acestea să fie complementare, chiar dacă modelele funcționează pe baza acelorași principii, fiind de același tip. De exemplu, în cazul modelelor HMM-GMM, criteriul de optimizare la antrenare poate fi de tip generativ sau discriminativ. Clasificatorii

generativi încearcă să învețe modelul care generează datele, calculând probabilitățile și distribuțiile modelului. În cazul clasificatorilor discriminativi, probabilitățile posterioare ale stărilor sunt modelate direct, astfel încât discriminarea între foneme se face cât mai bine. În (Shen et al., 2016), se folosesc atât modele SGMMM, ce sunt de tip generativ, cât și FBMMI, de tip discriminativ.

Algoritmii de decodare ai semnalului vocal sunt un alt factor ce conduce la obținerea de sisteme complementare. În (Jouvet and Fohr, 2014) decodarea este realizată atât în ordine normală, cât și în ordine inversă. Pentru cel de-al doilea caz, s-a recurs la inversarea cadrelor de semnal de la intrarea sistemului, împreună cu inversarea cuvintelor din modelul de limba și din modelul fonetic. Un element ce poate oferi transcrieri complementare în faza de decodare este modelul de limbă aplicat. Decodarea pot fi de asemenea directă, folosind un model de limbă pentru a obține graful cu transcrieri alternative (latice) sau decodare urmată de reevaluarea laticei, folosind tehnica numită *rescoring* (Bell et al., 2013).

3 Îmbinarea și aplicarea metodelor de complementaritate

În articolul (Lojka and Juhár, 2014) se utilizează diversitatea la nivelul trăsăturilor acustice, tipurilor de vorbitori (masculini/feminini) și a seturilor de date de antrenare pentru modelul de limbă. Se folosesc mai multe sisteme RAV pentru limba slovacă, antrenate pe aceleași baze de date, folosind aceeași paradigmă, platforma HMM-GMM. Fiecare model conține 3 stări pentru HMM și 32 densități gaussiene. Complementaritatea este dată de tipul trăsăturilor vocale utilizate, cât și de datele cu care au fost antrenate modelele de limbă. Se folosesc coeficienți MFCC și PLP, împreună cu derivatele lor de ordin 1 și 2, iar în final a fost aplicată o metodă de normare a mediei acestora. Trei modele acustice au fost antrenate folosind trăsături de tip MFCC, câte unul pentru fiecare gen (masculin, feminin) și unul general. În schimb, un singur model acustic a fost antrenat folosind PLP, folosind întreg setul de date de antrenare. Două modele de limbă au fost utilizate, ambele fiind de tip n-gram. Unul dintre ele a fost antrenat cu text provenind dintr-un domeniu specific, în timp ce al doilea a fost antrenat cu text general. Sistemele RAV finale au fost obținute prin combinarea fiecărui model acustic cu fiecare

model de limbă, fiind obținute în total 8 sisteme.

(Garau and Renals, 2008) combină mai multe tipuri de trăsături pentru a obține sisteme RAV complementare. Modelele acustice au fost antrenate folosind platforma HMM-GMM. Modelul de bază a fost obținut folosind coeficienți MFCC, împreună cu derivatele de ordin 1 și 2. Câte 16 mixturi gaussiene au fost utilizate pentru a modela stările acustice. Alte modele au folosit coeficienți MFCC bazați pe reprezentarea STRAIGHT (Kawahara et al., 1999), combinări între vectorii cu coeficienți MFCC standard și MFCC STRAIGHT, coeficienți de tip PLP, coeficienți PLP bazați pe reprezentarea STRAIGHT, combinări între vectorii cu coeficienți PLP standard și PLP STRAIGHT. Alte variațiuni cuprind acești coeficienți în combinație cu aplicarea tehnicii VTLN.

În (Ma et al., 2012) complementaritatea constă în utilizarea unor algoritmi diferiți pentru antrenarea modelelor, împreună cu varierea numărului total de stări acustice. Toate modelele acustice sunt modelate cu HMM-uri cu 3 stări, dependente de context, fiecare stare fiind modelată de 16 densități gaussiene. Se pleacă de la două sisteme inițiale, ale căror modele acustice sunt de tip GMM, SGMM și SGMM inițializat cu parametri deja antrenați într-un model UBM. Modelele inițiale sunt antrenate cu puține date, ele fiind reantrenate prin adăugarea la corpusul de antrenare a datelor selectate. Trăsăturile acustice extrase sunt de tip PLP, împreună cu derivatele lor de ordin 1 și 2.

(Li et al., 2016) utilizează sisteme complementare atât din punct de vedere al tipului modelelor, cât și din punct de vedere al trăsăturilor acustice. Este antrenat un model acustic de tip HMM-GMM, ce folosește vectori de trăsături de tip PLP, 39-dimensionali, asupra cărora se aplică tehnici de normare a mediei (CMN) și variației (CVN). Modelul de tip DNN-HMM folosește trăsături de tip *filterbank* împreună cu derivatele lor, 40-dimensionale, luând în calcul câte 5 cadre de semnal la stânga și la dreapta cadrului curent. Stratul de intrare conține 1320 neuroni, stratul de ieșire conține 3000 neuroni, iar straturile ascunse, în număr de 7, conțin 1024 neuroni fiecare.

În (Koctúr et al., 2016) și (Koctúr et al., 2017) sunt folosite două sisteme RAV, cu modele acustice de tip HMM-GMM, complementaritatea lor fiind asigurată de seturile diferite de date cu care au fost antrenate.

În (Cucu et al., 2014) diversitatea celor două modele acustice este dată de modul de împărțire al setului de date de antrenare, respectiv tipul vorbirii din aceste seturi. Astfel, unul dintre sisteme a fost creat folosind un subset cu vorbire citită, în condiții de liniște, în timp ce al doilea se bazează pe

vorbire spontană, uneori în condiții de zgomot.

(Jouvet and Fohr, 2014) folosește ca sursă de diversitate doi algoritmi diferiți de decodare, aplicați în cadrul aceluiași sistem RAV. Altfel, decodarea semnalului vocal are loc în ordine normală și inversă. Modelul acustic a fost antrenat pe baza platformei HMM-GMM, folosind coeficienți MFCC. S-a utilizat același model fonetic și același model lingvistic, dar pentru situația când decodarea are loc în ordine inversă, aceste modele au fost inversate și ele, fiind scrise invers.

În (Shen et al., 2016) diversitatea se manifestă la nivelul tipului trăsăturilor extrase din semnalul vocal, tipul modelelor acustice și tipul algoritmilor de antrenare. Au fost utilizate 4 tipuri de trăsături acustice: coeficienți mel-cepstrali (MFCC), cepstrul liniar predictiv (PLP), coeficienți extrași cu ajutorul bancurilor logaritmice de filtre Mel (FBANK) și coeficienți (FBANK) împreună cu trăsături tonale. Din punctul de vedere al modelelor acustice, au fost utilizate atât modele bazate pe HMM-GMM, cât și pe rețele neuronale de tip DNN sau CNN-DNN. Algoritmii de antrenare folosiți au fost atât generativi, SGMM, cât și discriminativi, FBMMI. Pentru antrenarea acestora s-au utilizat alinierile date de un model simplu, bazat pe trifoneme dependente de context, antrenat cu HMM-GMM. Modelarea de limbă s-a făcut atât probabilistic, folosind modele de tip n-gram, cât și pe bază de rețele neuronale recurente (RNN). Modelele n-gram au fost utilizate în două faze: un model general pentru faza de decodare și un model adaptat la domeniu pentru faza de rafinare a textului transcris.

Următoarele 4 articole au utilizat ca metodă de diversitate decodarea cu modele acustice antrenate pe alte limbi, diferite de cea a corpusului decodat. Pentru această sarcină, fonemele din fiecare limbă pentru care există model acustic, au fost asociate la fonemele limbii țintă. În (Vu et al., 2010) s-au folosit modele acustice în limbile bulgară, rusă, poloneză și croată pentru a decoda o bază de date de vorbire în limba cehă. Transcrierile astfel obținute au fost utilizate mai departe la antrenarea unui sistem RAV în limba cehă. (Vu et al., 2011) are de asemenea ca scop decodarea și crearea unui sistem RAV pentru limba cehă, folosind de data aceasta modele acustice pentru limbile engleză, franceză, germană și spaniolă. (Oparin et al., 2013) utilizează sisteme RAV pentru limbile engleză, franceză și rusă pentru a obține un model acustic în limba letonă. Au fost folosite aproximativ 800 ore de vorbire neadnotată din transmisiuni online. Modelul de limbă a fost antrenat folosind corpus de text adunat online, împreună cu transcrieri ale ședințelor din Parlament. S-au utilizat atât modele probabilistice, de tip n-gram, cât

și modele bazate pe rețele neuronale. În (Haitao et al., 2016) se dorește antrenarea nesupervizată a unui sistem RAV pentru limba japoneză, folosind la decodarea corpusului neadnotat, modele antrenate pe limbile mandarină, engleză și coreeană.

În (Gollan et al., 2007) complementaritatea sistemelor este realizată în special la nivelul trăsăturilor acustice. Cele 4 sisteme RAV utilizează coeficienți cepstrali, la care se adaugă diferite transformări: MFCC + CMVN + VTLN + LDA (1), + SAT + MPE (2), + MLLR (3) , + LM rescoring (4). Cel de-al patrulea sistem utilizează în plus față de cel precedent tehnica de *rescoring* pentru a îmbunătăți transcrierile inițiale. A fost antrenat un model acustic de tip HMM-GMM pe o bază de date de 85.7 ore de vorbire în limba engleză, adnotată manual. Un corpus de vorbire neadnotată, numit European Parliament Plenary Sessions (EPPS), ce are o durată de 180 ore, a fost transcris folosind sistemele complementare.

În (Huang et al., 2013) diversitatea sistemelor este dată de tipul modelelor acustice folosite, HMM-GMM, dar și HMM-DNN, împreună cu tipul algoritmului folosit la antrenarea modelului, generativ (*MLE - Maximum likelihood Estimation*) sau discriminativ (*fMPE+bMMI - boosted maximum mutual information*).

4 Metode de selecție

Selecția datelor este ultima etapă din cadrul unui proces de adnotare automată a corpusurilor de vorbire. În urma folosirii unor metode ce asigură complementaritatea, sunt create mai multe sisteme RAV cu care se transcrie vorbirea neadnotată, obținându-se mai multe transcrieri alternative. Scopul selecției este de a combina toate informațiile disponibile pentru a rezulta în final cea mai bună transcriere pentru un segment audio. Se urmăresc două aspecte: o precizie cât mai bună, astfel încât transcrierea stabilită să fie cea corectă și obținerea de transcrieri pentru o cantitate cât mai mare din totalul datelor audio neadnotate. De asemenea, în funcție de caz, există un compromis între aceste aspecte, astfel încât se dorește o transcriere pentru o parte din corpus său pentru întreg corpusul, respectiv o încredere cât mai mare sau o încredere mai mică pentru transcrierea obținută.

Alinierea completă a două sau mai multe ipoteze obținute în urma procesului de decodare cu sisteme RAV reprezintă o primă și cea mai simplă metodă de selecție. Această aliniere presupune identificarea segmentelor

identice și considerarea lor ca fiind corecte, având în vedere premisa că sistemele complementare fac erori diferite. În (Cucu et al., 2014) selecția se face în acest fel, utilizând algoritmul Dynamic Time Warping (DTW). De asemenea, metoda este întâlnită în (Koctúr et al., 2016) și în (Koctúr et al., 2017).

Scorul de încredere al cuvintelor permite filtrarea părților selectate ca fiind comune în urma alinierii. Aceasta este o metodă similară cu metoda alinierii complete, dar mai eficientă. De obicei, sistemele RAV, scot la ieșire un graf de transcrieri alternative, denumit latice, unde fiecare nod este un cuvânt, iar fiecare arc este o tranziție între stări, cu o anumită probabilitate. Transcrierea finală este cea mai bună cale prin acest graf, calea cu cele mai mari probabilități, acestea purtând numele de scor de încredere. Un prag foarte jos pentru scorurile de încredere va permite acceptarea unor transcrieri eronate, în timp ce un prag foarte ridicat este posibil să ignore cuvinte ce sunt corecte. O astfel de abordare este întâlnită în (Ma et al., 2012), iar în (Koctúr et al., 2016) este aplicată o constrângere, ce vizează numărul minim de cuvinte dintr-o secvență selectată. Similar se regăsește și în (Jouvet and Fohr, 2014), unde pe lângă constrângerea precedentă, se cere și o durată minimă de non-vorbire între secvențele selectate. (Gollan et al., 2007) folosește aliniere și prăguirea scorului de încredere la nivel de cuvânt și la nivel de stare acustică.

ROVER (Fiscus, 1997) (Recognizer Output Voting Error Reduction) este un sistem ce are ca scop combinarea mai multor ipoteze rezultate în urma procesului de transcriere, astfel încât se obține o singură ipoteză, cât mai corectă. ROVER utilizează programarea dinamică pentru a construi un graf de tranziții între cuvinte (WTN - word transition network). Mai întâi, se selectează prima ipoteză și se consideră a fi ipoteză de bază. Apoi, se aliniază cu această cea de-a doua ipoteză, utilizând programarea dinamică și distanța Levenshtein. În urma alinierii, se obține o secvență ce conține cuvintele comune, regăsite în ambele ipoteze, dar sunt marcate și pozițiile în care au apărut inserții sau ștergeri de cuvinte. Această secvență este aliniată mai departe cu următoarea ipoteză și procesul se repetă până la final. Rezultatul obținut în urma ultimei alinieri este reprezentat sub forma unei rețele de confuzie a cuvintelor, un graf în care trecerea de la un cuvânt la următorul este formată din mai multe arce reprezentând alternative de cuvinte. În cazul în care două cuvinte diferite primesc același vot, ROVER da prioritate primei alternative din graf. Procesul de votare este dat de formula:

$$\text{ScorVot}(w) = \alpha N(w,i) + (1-\alpha)C(w,i),$$

unde $N(w,i)$ este numărul de apariții ale cuvântului w în setul de transcrieri alternative, $C(w,i)$ este scorul de încredere al cuvântului w, iar α este un parametru de pondere între frecvența de apariție a cuvintelor și scorurile lor de încredere. ROVER folosește trei scheme de vot ce se bazează pe:

- frecvența aparițiilor - ignoră informațiile despre scorul de încredere ($\alpha = 1$);
- frecvența de apariție și media scorurilor de încredere - calculează o medie a scorurilor de încredere pentru fiecare cuvânt;
- frecvența de apariție și maximul scorului de încredere - similar cu schema precedentă, doar că este luat în calcul cuvântul cu cel mai mare scor de încredere.

În (Lojka and Juhár, 2014) se întâlnește această abordare de selecție a datelor ce utilizează ROVER sub mai multe înfățișări. Este folosită frecvența de apariție a cuvintelor în transcrieri, dar și scorul de încredere. (Garau and Renals, 2008), (Bell et al., 2013) și (Huang et al., 2013) utilizează de asemenea metoda ROVER pentru a alinia ipotezele obținute de la mai multe sisteme RAV. În (Shen et al., 2016) se folosește ROVER, iar asupra scorurilor de încredere au fost aplicate mai multe praguri.

Stabilitatea acustică (A-stabil) este o altă metodă de selecție întâlnită în literatura de specialitate. Prin varierea ponderilor modelului acustic, cât și a modelului de limbă, se obțin mai multe transcrieri alternative pentru fiecare sistem. Se alege drept referință, o transcriere pentru care ponderea între modelul acustic și cel lingvistic este cea mai echilibrată. Se calculează frecvențele de apariție ale fiecărui cuvânt din referință în celelalte transcrieri alternative și se normează la numărul de transcrieri alternative. Pentru fiecare propoziție din referință, scorul A-stab este definit ca media scorurilor pentru fiecare cuvânt. Pe baza acestui scor se face selecția datelor. Această metodă este întâlnită în (Vu et al., 2010), (Vu et al., 2011) și (Haitao et al., 2016), unde complementaritatea este dată de folosirea unor modele acustice antrenate pe limbi diferite, folosind mai departe tehnica numită *cross-language transfer*, pentru a decoda vorbire dintr-o limbă țintă diferită.

Tehnicile de învățare automată sunt prezente în selecția datelor, utilizând clasificatori de tip *conditional random fields* (CRF). Clasificatorul selector este antrenat pentru a decide care ipoteză este corectă (sau mai bună)

dintre cele rezultate din transcrierea cu sistemele complementare. Clasificatorul verificator determină dacă ipoteza selectată este corectă sau nu. În urma acestui proces, se pot obține câteva categorii de aliniere:

- cele două ipoteze corespund și sunt corecte ambele;
- cele două ipoteze nu corespund, dar una dintre ele este corectă;
- cele două ipoteze corespund, dar ambele sunt eronate;
- cele două ipoteze nu corespund, iar niciuna dintre ele nu este corectă.

Scopul clasificatorului este de a accepta datele utile (corespunzătoare categoriilor 1 și 2) și de a elimina datele eronate (categoriile 3 și 4). Metoda a fost utilizată în (Li et al., 2016).

Rețelele neuronale sunt de asemenea utilizate în sarcina de selecție a datelor. În (Koctúr et al., 2017), o metodă derivată din metoda alinierii simple, pleacă de la premisa că nu întotdeauna ipotezele au aceeași lungime, intervenind posibilitatea ca ipoteze scurte să fie aliniate cu ipoteze lungi. Astfel, se întâmplă uneori că unele cuvinte să fie aliniate la o altă parte din ipoteză, decât cea din care face parte. Acest inconvenient se rezolvă prin limitele temporale obținute de la sistemul RAV în etapa de decodare. În acest fel, se consideră incorect aliniate cuvintele ai căror indici temporali sunt foarte depărtați. În urma alinierii, se selectează secvențe de cuvinte consecutive de lungime n (n-grame), ce sunt verificate mai departe de către o rețea neuronală, pentru a stabili dacă secvența respectivă este corectă sau nu.

5 Îmbinarea și aplicarea metodelor de selecție

În (Lojka and Juhár, 2014) este folosit sistemul ROVER în câteva configurații. Prima dată este luat în calcul numai numărul de apariții ale cuvintelor în transcrierile alternative, apoi acesta este considerat împreună cu probabilitățile posterioare. A treia situație ține cont de numărul de apariții și de scorul de încredere al cuvintelor. La final, se reconsideră cazurile precedente, aplicându-se în plus un proces de netezire a scorurilor.

În (Huang et al., 2013) trei sisteme RAV ce diferă prin tipul modelelor acustice și prin modelele de limba sunt folosite pentru a genera ipoteze inițiale, cu scor de încredere la nivel de cuvânt și frază. Combinarea ipotezelor se face

folosind un sistem ROVER, ce duce la obținerea unor ipoteze mai bune și a unor noi scoruri de încredere. Acestea sunt obținute prin interpolarea liniară a scorurilor inițiale și a gradului de acord asupra ipotezei. Apoi, o procedură de votare ce cuprinde un sistem principal (obținut în urma combinării cu ROVER) și trei sisteme suplimentare (sistemele inițiale) sunt utilizate pentru a recalibra scorurile de încredere. Se va considera, pe rând, că fiecare sistem este cel principal. Numai ipoteza care provine de la sistemul principal este considerată ca fiind potențial corectă, în timp ce sistemele suplimentare au rolul de a calibra scorul de încredere al sistemului principal.

În (Shen et al., 2016) reducerea erorilor de transcriere prin aplicarea metodei ROVER a condus la îmbunătățirea performanțelor. Transcrierile obținute de la diferite sisteme au fost combinate, selecția datelor făcându-se pe baza unui scor de încredere. S-a testat aplicarea mai multor praguri asupra acestor scoruri.

În (Li et al., 2016) un set de clasificatori de tip *conditional random fields (CRF)* sunt folosiți pentru a selecta și verifica cea mai bună ipoteză dintre cele rezultate din transcrierea cu sistemul GMM-HMM sau DNN-HMM la nivel de caracter/cuvânt. Se folosește un clasificator selector, care alege una dintre ipoteze, împreună cu un clasificator verificator care determină corectitudinea ipotezei selectate. Deoarece se observa un tipar diferit între ipotezele obținute cu GMM față de DNN, clasificatorul CRF combină această diversitate și determină ce ipoteză ar trebui selectată pentru a reantrena modelul acustic. La început, textele sunt aliniate în perechi la nivel de caracter. O ipoteză corectă sau îmbunătățită este selectată pe baza ipotezelor complementare. În urma acestui proces, se pot obține 5 categorii de aliniere: cele două ipoteze corespund și sunt corecte ambele (1), cele două ipoteze corespund, dar ambele sunt eronate (2), cele două ipoteze nu corespund, iar niciuna dintre ele nu este corectă (3) sau cea bazată pe DNN este corectă (4) sau cea bazată pe GMM este corectă (5). Setul de clasificatori utilizează mai multe tipuri de trăsături, separate în două mari clase: trăsături bazate pe sistemul RAV (scorul de încredere al cuvântului, durata cuvântului, scorul cuvântului dat de modelul de limbă, scorul cuvântului dat de modelul acustic, numărul de cuvinte din latice conectate la cuvântul curent, numărul de cuvinte suprapuse în latice peste cuvântul actual) și trăsături bazate pe text (partea de vorbire reprezentată de cuvânt, probabilitatea dată de un model de limba 5-gram, comportamentul de tip back-off rezultat dintr-un model de limba 5-gram).

(Ma et al., 2012) realizează selecția transcrierilor obținute pe baza scorurilor de încredere. Se caută un prag al acestui scor, punând în balanță calitatea

datelor obținute și cantitatea lor. S-a observat că scorul de încredere la nivel de propoziție este mai relevant decât scorul la nivel de cuvânt.

În (Koctúr et al., 2016) ipotezele sunt aliniate complet și se consideră că părțile comune nu conțin erori, deși în realitate acestea au apărut, dar într-un procent scăzut. Mai departe, s-a încercat scăderea erorii pe baza scorului de încredere. Au fost testate mai multe praguri, iar cu cât pragul a fost mai înalt, cu atât eroarea a fost mai mică. De asemenea, au fost propuse câteva metode de normare a pragului și s-a încercat stabilirea unor formule pentru a obține rezultate mai bune. Combinația a două constrângeri, minimul de cuvinte consecutive și utilizarea unui prag pentru scorul de încredere, a dus la o scădere a erorii. Cel mai mare câștig a fost obținut prin utilizarea pragului variabil.

În (Cucu et al., 2014) se realizează alinierea ipotezelor și selecția părților comune. Aceeași abordare se întâlnește în (Jouvet and Fohr, 2014), iar mai departe se folosesc informațiile date de scorul de încredere și se aplică diferite praguri, pentru a efectua selecția finală.

În (Koctúr et al., 2017) ipotezele sunt aliniate, fiind stocate informații despre limitele temporale, forma ortografică, scorul dat de modelul acustic și scorul de încredere. Selecția ipotezelor se face cu ajutorul unui clasificator bazat pe rețele neuronale, ce încearcă să stabilească un model pentru secvențele de cuvinte consecutive (clasificarea n-gramelor). Acesta este mai eficient decât stabilirea unui prag pentru scorul de încredere, deoarece această metodă din urmă omite cuvintele cu un scor mic. Totodată, există cazuri când cuvinte transcrise incorect au un scor de încredere înalt, sarcina de decizie fiind dificilă. Pentru a antrena rețeaua neuronală, este necesar un corpus de date pentru care există informații despre corectitudinea fiecărui n-gram existent. Este important ca datele de antrenare să fie echilibrate, având aceeași cantitate de n-grame corecte și incorecte. Ordinul n-gramelor este cuprins între 1 și 6. Într-un prim experiment, vectorul de trăsături de la intrarea rețelei a fost compus din scorul dat de modelul acustic, împreună cu scorul de încredere, obținute de la cele două sisteme RAV, pentru fiecare cuvânt. Al doilea experiment a exclus scorul dat de modelul acustic. Al treilea experiment este similar cu al doilea, diferența fiind reprezentată de scalarea scorului de încredere prin înmulțirea lui cu un factor constant. Cel de-al patrulea experiment a folosit un singur scor de încredere, obținut ca o medie a celor furnizate de cele două sisteme RAV.

În (Gollan et al., 2007) selecția ipotezelor în scopul adnotării automate a fost realizată folosind două criterii: aliniere și stabilirea unui prag pentru

scorul de încredere la nivel de cuvânt, precum și alinierea și stabilirea unui prag pentru scorul de încredere la nivelul stărilor acustice.

În (Vu et al., 2010) selecția datelor se face pe baza scorurilor de încredere, calculate fie cu algoritmul *gamma*, fie pe baza stabilității acustice (A-stabil). Prima metodă presupune calculul probabilității unei căi din latice, în același fel în care algoritmul forward-backward funcționează la decodarea HMM-urilor. Un nod în latice poate fi văzut că o stare HMM, iar o cale din latice este o posibilă tranziție între stări. Deoarece nodurile sunt asociate cuvintelor din ipoteză, probabilitatea de emisie a unui nod este scorul acustic pentru acel segment temporal. Probabilitatea de tranziție între cuvinte este dată de modelul de limbă. Aceste două probabilități, date de modelul acustic și de modelul de limbă, compun scorul de încredere, puternic corelat cu eroarea de recunoaștere. Metoda stabilității acustice presupune generarea mai multor ipoteze. Fiecare ipoteză este aliniată cu o ieșire de referință, definită ca ipoteza cu cea mai bună pondere între modelul acustic și cel de limbă. Pentru fiecare cuvânt din referință, scorul de încredere este definit ca numărul de apariții ale cuvântului în ipotezele alternative, raportat la numărul de ipoteze alternative.

6 Rezultate experimentale

Experimentele ce au utilizat metodele de diversitate și selecție prezentate sunt destul de eterogene, iar o comparație directă între acestea este dificil de realizat. Totuși, unii autori au făcut specificații cu privire la performanțele unor metode, în comparație cu altele. De asemenea, au fost oferite unele date numerice cu privire la performanțele obținute, astfel încât se poate obține o privire de ansamblu asupra eficienței metodelor.

Tabelul 1 sumarizează metodele discutate în această lucrare. Pentru fiecare dintre acestea, se specifică metoda de complementaritate utilizată, numărul sistemelor complementare, metoda de selecție folosită, informații despre cantitatea de date selectate și acuratețea lor, precum și îmbunătățirea sistemelor după reantrenarea cu noile date obținute.

Articol	Complementaritate	#	Selecție	Date selectate	WER	Date selectate (h) / total date (h)	Îmbunătățire relativă
Lojka and Juhár (2014)	trăsături și modele de limbă	8	ROVER	–	–	–	–
Garau and Renals (2008)	combinații de trăsături spectrale	5–7	ROVER	–	–	–	3.2–7%
Ma et al. (2012)	modele acustice	3 (sau 8)	scor la nivel de latice	30%	25%	9 / 30	1.5–3%[c]
				76%	42%	23 / 30	
Li et al. (2016)	modele acustice și trăsături	2	clasificator CRF	68%	21%[a]	78.3 / 114.7	7%
Koctúr et al. (2016)	seturi de antrenare	2	alinierea ipotezelor + scor	0.3%	6%	–	56%
Koctúr et al. (2017)	seturi de antrenare	2	alinierea n-gramelor (ordin 1–6) + verificarea cu o rețea neuronală	–	16%[b]	–	–
Jouvet and Fohr (2014)	decodări în sens normal și invers	2	alinierea ipotezelor	39%	–	542 / 1377	0.28%
			alinierea ipotezelor + scor (prag 0.4)	42%	–	585 / 1377	0.51%
			alinierea ipotezelor + scor (prag 0.6)	19%	–	270 / 1377	1.27%
Shen et al. (2016)	modele acustice și trăsături	–	ROVER + scor (prag 0.1)	100%	7.3%	–	17–21%[d]
			ROVER + scor (prag 0.9)	86%	8%	–	
Vu et al. (2010)	modele acustice pentru limbi diferite	4	stabilitate acustică (A-stab)	81%	14.5%	18.5 / 23	
Vu et al. (2011)	modele acustice pentru limbi diferite	4	stabilitate acustică (A-stab)	72%	14.5%	–	–
Oparin et al. (2013)	modele acustice pentru limbi diferite	3	–	97%	–	783 / 800	–
Haitao et al. (2016)	sisteme RAV pentru limbi diferite	3	stabilitate acustică (A-stab)	20%	51.7%[a]	20 / 100	24%
Gollan et al. (2007)	modele acustice cu trăsături diferite + aplicarea tehnicii de rescoring	4	scor la nivelul cuvintelor + scor la nivelul stărilor acustice	80%		146 / 182	7%
Cucu et al. (2014)	seturi de antrenare, tipul vorbirii	2	DTW	–	9%-10%	–	–
Bell et al. (2013)	trăsături extrase cu DNN	2	ROVER	49%	–	98 / 200	5%
Huang et al. (2013)	modele acustice, și modele de limbă	3	ROVER	60%	–	–	7%-15%

[a] se referă la eroarea de la nivel de caracter (ChER)
[b] se referă la eroarea de clasificare a n-gramelor
[c] în funcție de tipul vorbirii din seturile de evaluare
[d] în funcție de setul de evaluare

Tabelul 1: Sumar al metodelor.

În Ma et al. (2012) se face o apreciere asupra scorurilor de încredere, observându-se că acestea sunt mai relevante la nivel de propoziție decât la nivel de cuvânt.

(Koctúr et al., 2017) folosește o rețea neuronală pentru a stabili corectitudinea secvențelor n-gram selectate din transcrieri. Această metodă a obținut o acuratețe de 84.55%. Utilizarea scorului dat de modelul acustic a condus la rezultate mai slabe față de situațiile în care acesta a fost exclus. De asemenea, s-a observat că este mai util ca la intrarea rețelei neuronale să existe două valori ale scorului de încredere, în locul uneia singure.

(Jouvet and Fohr, 2014) observă că în experimente anterioare, părțile comune din cele două transcrieri sunt corecte în proporție de aproximativ 90%. Transcrierile au fost obținute prin decodarea normală și inversă, folosind același sistem RAV. Două constrângeri suplimentare au fost adăugate la selecția datelor: segmentele trebuie să fie de cel puțin 10 cuvinte și să fie precedate sau succedate de cel puțin 300 ms de non-vorbire. Reantrenarea sistemului cu datele inițiale împreună cu datele nou obținute, a condus la o scădere absolută cu 0.4% a erorii la nivel de cuvânt. Metoda selecției pe baza transcrierilor obținute prin decodare în ordine normală și inversă a surclasat selecția datelor pe baza scorului de încredere.

În (Shen et al., 2016) rezultatele au arătat că pentru un prag jos al scorului de încredere, cuprins între 0.1-0.3, procentul datelor selectate este de aproape 100%, obținându-se pentru acestea un WER de aproximativ 7%.

Gollan et al. (2007) reușește ca prin aplicarea complementarității la nivelul modelării acustice să obțină o selecție a datelor în proporție de 80%, iar îmbunătățirea relativă adusă sistemului RAV final este de 7%.

În (Li et al., 2016), selecția și verificarea ipotezelor utilizând clasificatori de tip *conditional random fields* surclasează alte încercări bazate pe scor de încredere sau combinarea sistemelor folosind ROVER.

7 Concluzii

Ideea utilizării sistemelor RAV complementare este una fiabilă în contextul adnotării automate a corpusurilor de vorbire. Diversitatea sistemelor poate fi dată de mai mulți factori, cei mai importanți fiind compoziția seturilor de antrenare acustică/lingvistică, tipul trăsăturilor extrase, tipul modelelor antrenate, tipul algoritmilor utilizați la antrenare/decodare. Nu se poate afirma cu certitudine că vreuna dintre metode oferă complementaritate mai

bună, fiecare poate fi eficientă într-o situație particulară.

În ceea ce privește metodele de combinare și selecție, se observă faptul că ROVER este o procedură clasică, ce a fost aplicată cu succes în multe situații. O altă metodă clasică este cea a utilizării scorurilor de încredere, deși nu întotdeauna sistemele RAV scot la ieșire scoruri corelate cu acuratețea transcrierii. Totuși, alternativele ce se bazează pe tehnici de învățare automată (clasificatori CRF, rețele neuronale) au început să fie utilizate în ultima vreme, iar rezultatele sunt mai bune în comparație cu metodele clasice. Metodă stabilității acustice este derivată din ROVER și se folosește preponderent în situațiile în care se folosesc sisteme în limbi sursa diferite față de limba sursă.

Bibliografie

Bell, P., Yamamoto, H., Swietojanski, P., Wu, Y., McInnes, F., Hori, C., and Renals, S. (2013). A lecture transcription system combining neural network acoustic and language models. In *INTERSPEECH*, pages 3087–3091.

Cucu, H., Buzo, A., Besacier, L., and Burileanu, C. (2015). Enhancing asr systems for under-resourced languages through a novel unsupervised acoustic model training technique. *Advances in Electrical and Computer Engineering*, 15(1):63–68.

Cucu, H., Buzo, A., and Burileanu, C. (2014). Unsupervised acoustic model training using multiple seed asr systems. In *Spoken Language Technologies for Under-Resourced Languages*.

Fiscus, J. G. (1997). A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER). In *Workshop on Automatic Speech Recognition and Understanding*, pages 347–354. IEEE.

Garau, G. and Renals, S. (2008). Combining spectral representations for large-vocabulary continuous speech recognition. *Transactions on Audio, Speech, and Language Processing*, 16(3):508–518.

Gollan, C., Hahn, S., Schlüter, R., and Ney, H. (2007). An improved method for unsupervised training of lvcsr systems. In *Eighth Annual Conference of the International Speech Communication Association*.

Haitao, Y., Ji, X., and Jian, L. (2016). Multi-lingual unsupervised acoustic modeling using multi-task deep neural network under mismatch conditions. In *2016 8th IEEE International Conference on Communication Software and Networks (ICCSN)*, pages 139–144.

Huang, Y., Yu, D., Gong, Y., and Liu, C. (2013). Semi-supervised GMM and DNN acoustic model training with multi-system combination and confidence re-calibration. In *Interspeech*, pages 2360–2364.

Jouvet, D. and Fohr, D. (2014). About combining forward and backward-based decoders for selecting data for unsupervised training of acoustic models. In *Interspeech*, pages 815–819.

Kawahara, H., Masuda-Katsuse, I., and De Cheveigne, A. (1999). Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds1. *Speech communication*, 27(3-4):187–207.

Koctúr, T., Ondáš, S., and Juhár, J. (2017). Speech corpus generation based on n-gram confidence measure classification. In *International Symposium ELMAR*, pages 149–152.

Koctúr, T., Staš, J., and Juhár, J. (2016). Unsupervised acoustic corpora building based on variable confidence measure thresholding. In *International Symposium ELMAR*, pages 31–34.

Li, S., Akita, Y., and Kawahara, T. (2016). Data selection from multiple ASR systems' hypotheses for unsupervised acoustic model training. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5875–5879.

Lojka, M. and Juhár, J. (2014). Hypothesis combination for slovak dictation speech recognition. In *International Symposium ELMAR*, pages 1–4. IEEE.

Ma, Z., Wang, X., and Xu, B. (2012). Unsupervised training of subspace Gaussian mixture models for conversational telephone speech recognition. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4829–4832.

Oparin, I., Lamel, L., and Gauvain, J. L. (2013). Rapid development of a latvian speech-to-text system. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7309–7313.

Shen, P., Lu, X., Hu, X., Kanda, N., Saiko, M., Hori, C., and Kawai, H. (2016). Combination of multiple acoustic models with unsupervised adaptation for lecture speech transcription. *Speech Communication*, 82:1 – 13.

Vu, N. T., Kraus, F., and Schultz, T. (2010). Multilingual a-stabil: A new confidence score for multilingual unsupervised training. In *2010 IEEE Spoken Language Technology Workshop*, pages 183–188.

Vu, N. T., Kraus, F., and Schultz, T. (2011). Cross-language bootstrapping based on completely unsupervised training using multilingual a-stabil. In

Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on, pages 5000–5003. IEEE.