

Raport de analiză a impactului utilizării scorurilor de încredere pentru filtrarea transcrierilor RAV în vederea reantrenării sistemelor RAV

*Dan Oneață, Alexandru Caranica,
Horia Cucu, Dragoș Burileanu, Corneliu Burileanu*

1. Introducere

În acest raport descriem soluția îmbunătățită pentru estimarea scorurilor de încredere. Scopul nostru este de a înzestra sistemele automate cu estimări cât mai fiabile a încrederii în predicțiile pe care acestea le generează. Ideal, cu cât sistemul este mai încrezător într-o predicție (adică un scor de încredere mai mare), cu atât ne așteptăm ca rezultatul prezis să fie corect. Alternativ, putem rezolva problema complementară a estimării incertitudinii — în acest caz, cu cât este mai mare scorul de incertitudine, cu atât este mai probabil ca predicția să fie eronată.

În contextul proiectului de față, și anume, al recunoașterii automate a vorbirii (RAV) și al adnotării automate a datelor audio, estimarea încrederii este de o importanță crucială pentru că ne semnalează corectitudinea datelor transcrise automat și ne permite o filtrare pe baza acestui scor. Dar metodele pentru estimarea scorurilor de încredere pentru RAV au aplicații multiple, care merg dincolo de sarcina noastră de interes: îmbunătățirea robusteții sistemelor în sarcini critice de siguranță, evitarea erorilor în sistemele de dialog om-mașină sau facilitarea corecțiilor manuale în sarcinile de transcriere audio prin semnalarea erorilor. Mai mult, lucrări anterioare de specialitate au valorificat estimările scorurilor de încredere pentru o serie de sarcini care depind de RAV: selectarea predicțiilor cu grad ridicat de încredere pentru reantrenarea sistemului de bază [Sperber, 2017], propagarea incertitudinilor în traducerea automată a vorbirii [Vesely, 2013], adnotarea manuală a predicțiilor mai puțin sigure pentru învățarea activă [Yu, 2010].

Metoda pe care o propunem vizează estimarea încrederii pentru sistemele RAV de la un capăt la altul (en., end-to-end) [Hadian, 2018] — spre deosebire de soluția de bază care consideră sistemele RAV de tip hibrid, DNN-HMM. Modelele RAV de tip end-to-end au câștigat masiv în popularitate în ultima perioadă nu doar datorită performanței lor (care o egalează și chiar depășește pe cea a RAV-urilor de tip clasic), dar și pentru avantajele suplimentare de a fi simple din punct de vedere conceptual și de a permite un proces de antrenare unitar [Lüscher, 2019; Tüske, 2019; Karita, 2019]. Cu toate acestea, surprinzător de puține lucrări tratează sarcina de estimare a încrederii în sistemele RAV de la un capăt la altul.

Metoda noastră adresează două provocări principale ale dezvoltării metodelor de estimare a încrederii pentru sistemele RAV: (i) ieșirea structurată (textul) și (ii) predicțiile granulare (la nivel de grafem sau token în loc de cuvinte). Sistemele RAV sunt modele structurate (care mapează secvențe la secvențe) spre deosebire de rețelele obișnuite de recunoaștere (cum ar fi clasificarea imaginilor) a căror ieșire este o singură etichetă. Natura secvențială a ieșirii impune o etapă de decodare, ceea ce complică nu numai predicția, cât și algoritmul de estimare a încrederii, deoarece acesta trebuie să opereze într-un context auto-regresiv (folosind secvența deja prezisă). Din acest motiv, obținem predicțiile pe baza unui RAV preantrenat și apoi aplicăm metodele de estimare a încrederii peste probabilitățile următorului simbol, care sunt condiționate de transcrierea fixă.

Pentru a evita constrângerile impuse de utilizarea unui dicționar de cuvinte fix și pentru a permite predicții de cuvinte noi, sistemele RAV de tip end-to-end de obicei folosesc la ieșire token-uri de subcuvinte. Dar având în vedere că token-urilor le lipsește semantica, pentru multe aplicații finale suntem interesați să estimăm încrederea la nivelul cuvintelor și nu a token-urilor. În acest scop, explorăm metode de agregare a măsurilor de incertitudine de la nivel de token la unități mai mari, corespunzând cuvintelor. Dar tehnicile prezentate în continuare pot fi aplicate și la structuri chiar mai ample, cum ar fi nivelul propoziției sau al enunțului.

Acest raport de analiză face referire în multiple rânduri la raportul științific și tehnic aferent proiectului component TADARAV [Georgescu, 2020] disponibil online pe pagina web a acestui proiect de cercetare¹.

2. Legături cu starea artei

În această secțiune facem o scurtă prezentare a lucrărilor de specialitate relevante pentru sarcina și metoda propusă. Considerăm două direcții: metode pentru scoruri de încredere pentru recunoașterea automată a vorbirii (RAV) și metode pentru scoruri de încredere pentru sisteme de tip end-to-end.

Scoruri de încredere pentru recunoașterea automată vorbirii. Cele mai multe lucrări anterioare privind scorurile de încredere pentru RAV vizează sisteme clasice, bazate pe paradigma HMM-GMM. Aceste metode extrag mai întâi un set de caracteristici din rețeaua de decodare, modelul acustic sau de limbă, și apoi antrenează un clasificator pentru a prezice dacă transcrierea este corectă sau nu. Exemple tipice de caracteristici includ probabilitatea realizării acustice, scorul modelului de limbă, durata cuvântului, numărul de alternative din rețeaua de confuzie [Kemp, 1997; Weintraub, 1997; Hazen, 2002]. Mai recent, Swarup *et al.* [Swarup, 2019] a mărit setul de caracteristici folosind embedding-uri profunde ale semnalului acustic de intrare și ale textului prezis, în timp ce Errattahi *et al.* [Errattahi, 2018] a arătat că adaptarea la domeniu a caracteristicilor extrase aduce beneficii de performanță. Clasificatorii folosiți de metodele de estimare a scorurilor de încredere variază de la conditional random fields [Seigel, 2013; Cortina, 2016] și perceptron cu mai multe straturi [Kalgaonkar, 2015] la rețele neuronale recurente bidirecționale [Ogawa, 2017; Del-Agua, 2018; Li, 2019].

Scoruri de încredere în sistemele end-to-end. Metoda de bază pentru estimarea încrederii în rețelele neuronale este de a utiliza în mod direct probabilitatea predicției celei mai probabile [Hendrycks, 2016]. S-a observat că rețelele neuronale tind să fie prea sigure și estimările probabilității pot fi îmbunătățite prin scalarea temperaturii [Hinton, 2015], ceea ce duce de obicei la o mai bună calibrare [Guo, 2017; Ashukha, 2020]. Cea mai promițătoare direcție în ceea ce privește simplitatea și utilitatea implică estimarea Monte Carlo: Gal și Ghahramani [Gal, 2016] folosesc tehnica dropout la momentul inferenței pentru a obține predicții multiple, care sunt apoi mediate, în timp ce Lakshminarayanan *et al.* [Lakshminarayanan, 2017] fac media predicțiilor unui ansamblu de rețele de obicei antrenate cu inițializări diferite. Această ultimă metodă s-a dovedit a fi foarte fiabilă pentru condițiile dificile în care avem date din afara domeniului [Ovadia, 2019], dar este costisitoare din punct de vedere computațional, deoarece implică antrenarea a multiple modele de RAV [Ashukha, 2020]. O abordare diferită a estimării încrederii este de a învăța un clasificator (de obicei o altă rețea neuronală) direct deasupra activărilor rețelei de RAV [Corbière, 2019; Chen, 2019].

La intersecția acestor două direcții de cercetare, menționăm lucrarea foarte recentă a lui Malinin și Gales [Malinin, 2020], care similar cu noi abordează sarcina de estimare a încrederii pentru sistemele RAV end-to-end. Cu toate acestea, ei sunt preocupați de estimarea incertitudinii la nivel de token și frază, în timp ce noi suntem interesați de estimarea la nivel de cuvânt, și, în consecință, oferim mai multă atenție asupra tehnicilor de agregare. Mai mult, ei folosesc ansambluri ca metodă principală de estimare a încrederii, în timp ce noi evaluăm și metodele de reducere a temperaturii și de scădere. Deși tehnica de dropout a fost folosită anterior

¹ Proiectul de cercetare TADARAV: <https://tadarav.speed.pub.ro>

pentru obținerea scorurilor de încredere pentru RAV [Vyas, 2019], metoda este diferită de abordarea noastră. În acea lucrare, autorii generează mai multe ipoteze prin dropout și apoi atribuie confidențe cuvintelor pe baza frecvenței aparițiilor lor în ipotezele aliniate. În schimb, noi agregăm probabilitățile posterioare și nu ipotezele, ceea ce simplifică procedura, deoarece evită pasul de aliniere.

3. Metodologia

În această secțiune prezentăm efectiv metodologia propusă pentru estimarea scorurilor de încredere și modalități propuse de îmbunătățire a acestora. Începem mai întâi cu o descriere generală a metodei și a notației implicate.

Considerăm un model de tip secvență-la-secvență (en. sequence-to-sequence) care mapează o secvență audio la una de token-uri. Modelul este specificat de parametri, care sunt învățați prin minimizarea pe setul de antrenare a unei funcții de pierdere, cum ar fi funcția *connectionist temporal classification* (CTC) sau divergența Kullback-Leibler. La inferență, modelul generează probabilități pentru următorul simbol într-o manieră autoregresivă, bazat pe token-urile deja prezise. Aceste probabilități sunt utilizate pentru efectuarea decodării prin metoda beam search pentru a obține cea mai probabilă secvență de token-uri. Având în vedere că probabilitatea de ieșire condiționată este o distribuție peste token-uri din vocabular, o notăm cu un vector \mathbf{p}_k -dimensional.

3.1 Estimarea scorurilor de încredere

Scopul nostru este de a obține un scor de încredere pentru fiecare cuvânt din transcrierea produsă de RAV. Realizăm acest lucru în doi pași. Mai întâi, folosind probabilitățile aposteriori de la fiecare moment de timp, extragem caracteristici pentru a reprezenta scorul de încredere al fiecărui token. În cea de-a doua etapă, agregăm scorurile la nivel de token în scoruri de încredere la nivel de cuvânt, pe baza token-urilor care aparțin fiecărui cuvânt. În continuare vom detalia acești doi pași; vezi și figura 1.

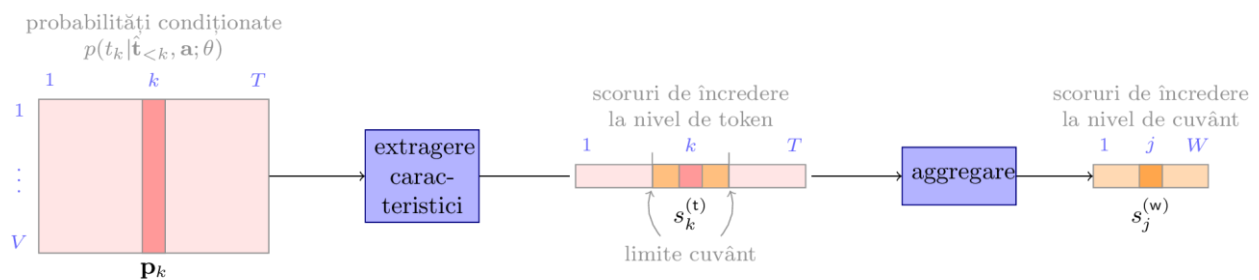


Figura 1. Prezentare schematică generală a metodei propuse de estimare a scorurilor de încredere. Pe baza unui sistem de recunoaștere a vorbirii (RAV) de tip end-to-end, obținem probabilități pentru fiecare token condiționate de o rostire și token-urile prezise anterior. Pe baza acestor probabilități extragem scoruri de încredere la nivel de token, pe care le agregăm apoi pentru a obține scoruri la nivel de cuvânt. Dimensiunea vocabularului de token-uri este notată cu V , numărul de token-uri este notat cu T și numărul de cuvinte cu W .

Extragerea caracteristicilor. Pentru a măsura încrederea într-o predicție la nivel de token folosim două variante:

- Logaritmul probabilității (**log-proba**) celei mai probabile predicții date de clasificator, adică $\log p_k$. S-a observat empiric că acest tip de caracteristică oferă un baseline puternic pentru sarcinile de clasificare a greșelilor și detectare a eșantioanelor din afara distribuției [Hendrycks, 2016].
- Entropie negativă (**neg-entropie**) calculată peste vocabularul token-urilor la fiecare moment de timp, adică $-\log \sum_{t \in \mathcal{V}} p(t | t_{<k}, a; \theta)$. O entropie mare înseamnă o incertitudine mare sau, invers, o entropie *negativă* mare implică o predicție încrezătoare. Deși entropia este de obicei utilizată ca măsură de încredere împreună cu tehnica de dropout [Gal, 2016], după cum vom vedea în curând, tehnica de dropout va putea fi folosită și peste probabilitățile originale.

Agregarea. Pentru a obține caracteristici la nivel de cuvânt din cele la nivel de simbol, experimentăm cu trei tipuri de funcții de agregare: suma, media, minimul. Deoarece ambele caracteristici propuse sunt negative, însumarea a mai multe token-uri va duce la valori mai mici și, prin urmare, la scoruri de încredere mai mici; acest comportament poate fi de dorit deoarece cuvintele mai lungi sunt mai susceptibile de a fi eronate. De

asemenea, atunci când însumăm logaritmul probabilităților, obținem un scor la nivel de cuvânt corespunzător probabilității log a întregii secvențe. Utilizarea agregării cu funcția de minim este justificată de faptul că am putea dori o încredere scăzută dacă cel puțin unul din token-uri are o încredere scăzută.

3.2 Îmbunătățirea probabilităților la nivel de token

Propunem trei moduri de a face mai fiabile probabilitățile la nivel de token: scalarea temperaturii, tehnica de dropout și ansamblurile de modele. Presupunerea noastră este că îmbunătățind probabilitățile la nivel de token, vom obține și scoruri de confidență la nivel de cuvânt mai bune.

Scalarea temperaturii [Hinton, 2015; Guo, 2017] constă în împărțirea activărilor de tip logit (valorile de dinainte de stratul softmax) la un scalar (cunoscut sub numele de temperatură). Valoarea lui variază de la zero la infinit și controlează forma distribuției: când obținem o distribuție uniformă, când obținem o distribuție Dirac localizată pe cea mai probabilă ieșire. Pe baza temperaturii actualizăm probabilitățile la nivel de token la fiecare moment de timp, după cum urmează:

Apoi extragem caracteristici peste probabilitățile actualizate, le agregăm în scorul la nivel de cuvânt și, în cele din urmă, clasificăm cuvântul drept corect sau incorect:

Variabilele sunt parametri și sunt învățate prin optimizarea unei funcții de pierdere de entropie încrucișată (en., *cross-entropy loss*) pe un set de validare. Etichetele sunt setate la nivel de cuvânt prin alinierea la textul de bază cu transcrierea. Remarcăm faptul că parametrii și nu modifică ordinea predicțiilor, dar ne permit să învățăm un model de estimare a încrederii calibrat.

Dropout [Srivastava, 2014] este o tehnică care maschează părți aleatorii ale activărilor într-o rețea, făcând rețeaua mai puțin predispusă la supra-antrenare (en., *overfitting*). În [Gal, 2016] s-a observat că tehnica de dropout induce o distribuție de probabilitate peste ponderile rețelei și poate fi în consecință utilizată pentru inferența Bayesiană aproximativă. Folosim această idee și calculăm probabilitățile la nivel de token obținute prin mai multe rulări folosind dropout:

$$\sum$$

unde specifică predicția dropout-ului. Probabilitățile actualizate sunt apoi utilizate pentru a extrage oricare dintre caracteristicile de incertitudine propuse (log-proba sau neg-entropie).

Ansamblurile [Lakshminarayanan, 2017] se bazează pe aceeași idee de mediere a predicțiilor din mai multe surse (ca în cazul dropout-ului), dar în acest caz ponderile provin din rețele antrenate independent (folosind inițializări diferite ale rețelei). În cazul nostru, calculăm media predicțiilor la nivel de token:

$$\sum$$

unde specifică ansamblul de modele. Este important de subliniat că trebuie să avem același context pentru toate modelele din ansamblu, deci îl folosim pe cel dat de un model preantrenat.

Cele trei abordări prezentate pot fi combinate; de exemplu, mai întâi putem actualiza probabilitățile folosind scalarea temperaturii și apoi media mai multe predicții folosind tehnica de dropout.

4. Setup-ul experimental pentru experimente pe limba engleză

În această secțiune descriem bazele de date, sistemele de recunoaștere automată a vorbirii utilizate și metricile de evaluare utilizate pentru evaluarea scorurilor de încredere pe limba engleză

4.1 Baze de date

Pentru experimentele pe limba engleză am ales mai multe seturi de date, toate disponibile public și utilizate de comunitate mai ales pentru evaluarea sistemelor de recunoaștere a vorbirii.

Tabelul 1 Dimensiunea seturilor de date (partea *test*) care sunt folosite pentru evaluarea metodelor pentru scorurile de încredere pe limba engleză.

Bază de date	Num. rostiri	Durață (ore)
Libri clean	2.6K	5.4
Libri other	2.9K	5.3
TED	1.1K	2.6
CommonVoice	66K	72

LibriSpeech [Panayotov, 2015] este un corpus de aproximativ 1000 de ore de cărți audio citite. Datele au fost derivate din proiectul LibriVox și au fost atent segmentate și aliniate. Folosim setul de date atât pentru antrenare, cât și pentru evaluare. Pentru antrenare folosim cele trei părți ale acestuia `clean100`, `clean360` și `other500`, în timp ce pentru validare și evaluare folosim părțile standard denumite `clean`, respectiv `other`.

TED-LIUM2 [Rousseau, 2014] constă în discursuri și transcrierile acestora colectate de pe site-ul web TED. Utilizăm setul de date pentru evaluare și, în consecință, folosim numai subseturile predefinite `dev` și `test`.

CommonVoice [Ardila, 2020] este un set de date colaborativ cu transcrieri scurte care sunt citite de oameni din întreaga lume. Există mai multe versiuni ale setului de date și am folosit prima versiune. Folosim setul de date pentru evaluare și am definit subseturi de `dev` și `test` prin alegerea a 10% din rostiri în mod aleatoriu pentru fiecare dintre ele.

Tabelul 1 prezintă dimensiunile părților de `test` pentru fiecare set de date de evaluare.

4.2 Sistemul de recunoaștere automată a vorbirii

Sistemul de recunoaștere automată a vorbirii (RAV) este implementat folosind librăria ESPnet [Watanabe, 2018]. Modelul implementează arhitectura de tip Transformer [Vaswani, 2017] și primește la intrare un banc de filtre Mel 80-dimensionale (extrase cu utilitarul Kaldi [Povey, 2011]) și produce la ieșire o secvență de token-uri. Vocabularul are dimensiunea de 5,000 de token-uri și este obținut prin segmentarea cuvintelor pe baza unui model de limbaj de tip unigram [Kudo, 2018]. Modelul este antrenat pe 960 de ore din baza de date LibriSpeech [Panayotov, 2015], iar datele sunt augmentate utilizând tehnicile SpecAugment (modificarea vitezei vorbirii, mascare în frecvență, mascare în timp) [Park, 2019]. Pentru decodare folosim un model de limbă, care este implementat tot ca Transformer și este antrenat pe 14,500 de cărți din domeniu public [Panayotov, 2015]. Vocabularul modelului de limbă constă din aceleași 5,000 de token-uri ca și modelul RAV.

Pentru experimentele cu ansambluri de modele, reantrenăm sistemul de RAV folosind aceeași arhitectură și date, dar inițializări ale ponderilor diferite. Repetăm procesul de patru ori obținând patru modele independente. Datorită constrângerilor de calcul, aceste modele au fost antrenate pentru un număr mai scurt de epoci decât sistemul principal (10 versus 120), dar am observat că curba funcției de pierdere a validării a început să se aplatizeze și că performanța sistemului de RAV este rezonabilă ($5.5\% \pm 0.4$ WER pe Libri `clean` față de 2.7% obținut de modelul preantrenat).

4.3 Metrice de evaluare

În mod ideal, ne dorim ca scorurile de încredere să fie corelate cu corectitudinea transcrierii, adică cuvintele corecte ar trebui să aibă un scor de încredere mare, în timp ce cuvintele incorecte, un scor scăzut. Pe baza lucrărilor anterioare [Hendrycks, 2016; Corbière, 2019; Malinin, 2020], folosim metrice care sunt utilizate în general pentru evaluarea clasificatorilor binari, dar variază pragul de discriminare între cele două clase. Mai precis, măsurăm aria de sub curba de precision-recall (area under precision-recall curve; AUPR) și aria de sub curba receiver operator curve (area under receiver operator curve; AUROC). Cu toate acestea, în funcție de ceea ce dorim să ne concentrăm (erori sau predicții corecte) obținem două variante: dacă suntem interesați de clasificări greșite, vom trata erorile ca pe o clasă pozitivă; pe de altă parte, dacă suntem interesați de clasificarea corectă, vom trata detecțiile reușite ca fiind clasa pozitivă. Prin urmare, pentru AUPR folosim două variante $AUPR_e$ (când erorile sunt tratate ca pozitive) și $AUPR_s$ (când succesele sunt tratate ca

pozitive). Pentru AUROC se obține aceeași valoare pentru ambele opțiuni, deci nu este necesar să se facă această distincție.

5. Rezultate experimentale pentru limba engleză

Această secțiune prezintă rezultatele experimentale pe bazele de date în limba engleză. Începem cu o evaluare a caracteristicilor și a agregărilor acestora; după care raportăm rezultatele pentru variantele îmbunătățite: mai întâi, pentru tehnicile de scalare a temperaturii și dropout, apoi pentru ansamblurile de modele. Încheiem secțiunea cu o discuție privind alte legături cu starea artei și posibile direcții.

5.1 Caracteristici și metode de agregare

Evaluăm caracteristicile pentru reprezentarea incertitudinii din token-uri și tehnicile de agregare propuse pe cele patru seturi de date descrise în secțiunea precedentă. Folosim modelul preantrenat pentru a obține predicții de text pentru toate fișierele audio din partea de testare al fiecărui set de date, și apoi estimăm încrederea pe baza metodologiei propuse. Tabelul 2 prezintă rezultatele pentru toate combinațiile de caracteristici și agregări.

Comparație a caracteristicilor. Observăm că prin folosirea caracteristicile de tip probabilități log se obțin rezultate mai bune față de situația utilizării caracteristicilor pe bază de entropie în toate cazurile (indiferent de agregare sau setul de date). Singura excepție notabilă este setul de date CommonVoice în care rezultatele sunt comparabile.

Tabelul 2. Rezultate pentru evaluarea scorurilor de încredere pentru toate combinațiile de caracteristici (caract.) și metode de agregare (agreg.) pe cele patru baze de date în engleză. Pentru toate cele trei metrici de evaluare utilizate (AUPRe, AUPRs, AUROC) valorile mai mari indică performanță mai bună. Raportăm eroarea la nivel de cuvânt pentru sistemul de RAV preantrenat pe fiecare din seturile de date; aceasta este listată în dreapta numelui bazei de date.

		Libri clean / 2.7%			Libri other / 6.0%		
caract.	agg.	AU-PR _e	AU-PR _s	AU-ROC	AU-PR _e	AU-PR _s	AU-ROC
log-proba	sum	21.55	99.21	82.41	29.99	98.10	81.75
log-proba	min	21.85	99.19	82.47	28.64	98.06	81.66
log-proba	avg	20.12	99.10	80.90	26.72	97.93	80.47
neg-entropy	sum	17.31	99.10	79.97	26.37	97.86	79.58
neg-entropy	min	19.94	99.09	80.55	26.75	97.82	79.64
neg-entropy	avg	17.55	98.95	77.72	24.26	97.59	77.46
		TED / 13.3%			CommonVoice / 28.6%		
caract.	agg.	AU-PR _e	AU-PR _s	AU-ROC	AU-PR _e	AU-PR _s	AU-ROC
log-proba	sum	39.97	95.88	79.95	48.98	77.71	64.84
log-proba	min	39.74	95.94	80.58	46.79	76.74	62.67
log-proba	avg	38.74	95.88	80.29	44.51	75.82	60.87
neg-entropy	sum	34.96	95.41	77.57	47.71	77.10	63.74
neg-entropy	min	37.55	95.56	79.01	45.51	76.00	61.21
neg-entropy	avg	36.28	95.42	78.29	42.64	74.83	58.75

Comparație a metodelor de agregare. În general, agregarea folosind suma funcționează mai bine cu caracteristicile de tip log-proba, în timp ce agregarea folosind minimul funcționează mai bine pentru caracteristicile de entropie. Suma s-ar putea să nu fie potrivită pentru caracteristicile de entropie, deoarece magnitudinea lor este mai mare decât în cazul caracteristicilor de tip log-proba, iar cuvintele lungi sunt

penalizate prea mult prin lungime; dar, așa cum vom vedea mai departe, acest comportament poate fi atenuat prin metoda de scalare a temperaturii. Media este, în general, slabă pentru ambele caracteristici, sugerând că măsurile invariante în lungime sunt dăunătoare. Într-adevăr, o privire mai atentă la frecvența erorilor cu dimensiunea lungimii indică următorul fapt: cu cât un cuvânt este format din mai multe token-uri, cu atât este mai probabil că este incorect, vezi figura 2.

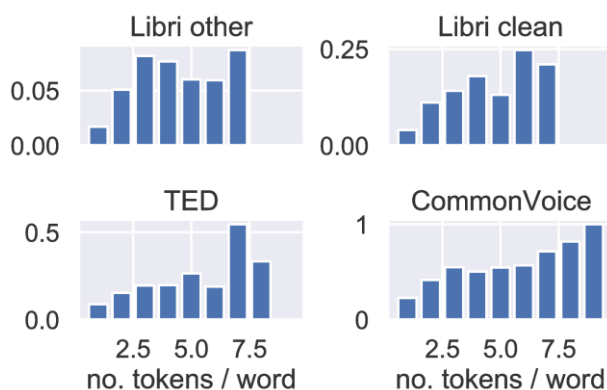


Figura 2. Raportul de erori ca o funcție de lungimea cuvântului. Raportul de erori este calculat ca numărul de cuvinte eronate împărțit la numărul total de cuvinte, iar lungimea cuvintelor este măsurată ca număr de token-uri.

Comparație între seturile de date. Așa cum era de așteptat, modelul preantrenat are cea mai bună performanță în ceea ce privește datele din domeniu (2.7% WER pe Libri clean și 6.0% pe Libri other), performanța scăzând apoi brusc, pe măsură ce evaluăm datele din afara domeniului (13.3% pe TED și 28.6% pe CommonVoice). În fiecare dintre aceste situații, numărul de cuvinte care sunt clasificate corect se schimbă, trecând de la mai multe pe baza de date Libri la mai puține pe seturile de date TED și CommonVoice. Această observație explică de ce performanța pentru AUPRs scade în funcție de domeniul datelor, și, dimpotrivă, de ce se îmbunătățește performanța AUPRe. Din păcate, din exact acest motiv—performanța diferită a sistemului RAV de bază pe cele patru seturi de date—este imposibil de comparat metodele de încredere între seturile de date, deoarece acestea utilizează o altă adnotare [Ashukha, 2020].

5.2 Scalarea temperaturii și tehnica de dropout

Evaluăm metodele de estimare a scorurilor de încredere după îmbunătățirea probabilităților token-urilor prin două din tehnicile descrise: scalarea temperaturii și tehnica de dropout. Folosim sistemul RAV preantrenat și raportăm rezultatele pe setul de testare TED. Parametrii pentru metoda de scalare a temperaturii sunt învățați pe partea dev a setului de date TED pentru fiecare setare de caracteristică și metodă de agregare. Când scalarea temperaturii este combinată cu tehnica de dropout, aplicăm mai întâi scalarea temperaturii (folosind aceeași temperatură) și apoi mediem probabilitățile obținute prin dropout. Pentru dropout mediem 64 de predicții independente. Tabelul 3 prezintă rezultatele pentru toate combinațiile de caracteristici, agregări și tehnici de îmbunătățire. Rezultatele indică faptul că ambele metode propuse îmbunătățesc rezultatele la fel ca și combinația lor, ceea ce oferă în general cel mai bun rezultat. Observăm că tehnica de dropout aduce îmbunătățiri mai mari pentru caracteristicile log-proba, în timp ce caracteristica neg-entropie produce rezultate mai bune atunci când se utilizează scalarea temperaturii. Interesant, cele mai bune rezultate sunt acum obținute pentru neg-entropie cu agregare sumă (rândul 16). Figura 3 arată că performanța dropout-ului se îmbunătățește cu numărul de rulări și se plafonează în jurul valorii alese de 64.

5.3 Ansambluri de modele

În cele ce urmează, prezentăm rezultate pentru estimarea scorurilor de încredere folosind ansambluri de modele și combinațiile acestora cu celelalte versiuni îmbunătățite (scalarea temperaturii și tehnica de dropout). Pentru fiecare dintre modelele reantrenate, care fac parte din ansamblu, folosim predicțiile modelului preantrenat pentru a selecta transcrierea pentru care dorim să estimăm scoruri de încredere; deci modelul reantrenat este folosit doar pentru scorul de încredere, prin extragerea caracteristicilor de încredere descrise anterior. Rezultatele sunt prezentate în tabelul 4. Pentru rândurile care nu folosesc ansamblu (rândurile 1, 2, 3 și 5 din tabel) evaluăm fiecare dintre cele patru modele individuale în mod independent și raportăm performanța medie. Modelul preantrenat (tabelul 3, rândul 13) are, în general, o performanță mai bună decât cele reantrenate (tabelul 4, rândul 1), sugerând că performanța predictivă a unui model se

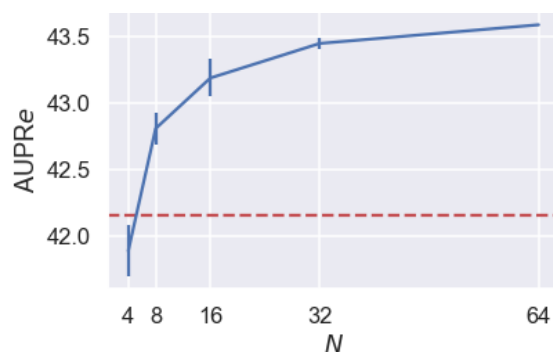


Figura 3. Performanța AUPRe ca funcție de numărul de rulări de dropout N pe baza de date TED. Linia roșie orizontală indică rezultatul metodei fără dropout. Modelul folosește caracteristici de tip neg-entropy, agregarea pe bază de sumă și scalare a temperaturii.

Tablelul 3. Evaluarea scorurilor de încredere pe baza de date TED pentru combinații de caracteristici, metode de agregare și tehnici de îmbunătățire a probabilităților: scalarea temperaturii (ST) și tehnica de dropout (D). Punctul negru “•” indică dacă e folosită o anumită tehnică, iar linia “–” indică opusul.

nr.	caract.	agreg.	ST	D	AUPRe	AUPRs	AUROC
1	log-proba	sum	–	–	39.97	95.88	79.95
2			–	•	41.41	96.81	82.78
3			•	–	40.92	96.19	81.11
4			•	•	42.99	97.14	84.10
5	log-proba	min	–	–	39.74	95.94	80.58
6			–	•	42.08	96.94	83.76
7			•	–	39.84	95.98	80.74
8			•	•	42.17	97.00	83.93
9	log-proba	avg	–	–	38.74	95.88	80.29
10			–	•	41.19	96.95	83.73
11			•	–	38.97	95.99	80.66
12			•	•	41.32	97.06	84.08
13	neg-entropy	sum	–	–	34.96	95.41	77.57
14			–	•	33.14	96.22	79.45
15			•	–	42.16	96.91	83.50
16			•	•	43.59	97.62	85.51
17	neg-entropy	min	–	–	37.55	95.56	79.01
18			–	•	38.75	96.53	81.98
19			•	–	41.23	96.87	83.50
20			•	•	42.23	97.60	85.51
21	neg-entropy	avg	–	–	36.28	95.42	78.29
22			–	•	38.01	96.51	81.85
23			•	–	40.22	96.53	82.48
24			•	•	41.15	97.43	85.18

Tabelul 4. Evaluarea scorurilor de încredere pe baza de date TED pentru combinații ale metodelor îmbunătățite: scalarea temperaturii (ST), dropout (D) și ansambluri (A). Am folosit caracteristica neg-entropy și suma ca agregare. Punctul negru “•” indică dacă e folosită o anumită tehnică, iar linia “–” indică opusul.

nr.	ST	D	A	AUPRe	AUPRs	AUROC
1	–	–	–	28.58	95.30	75.79
2	•	–	–	32.00	96.32	79.47
3	–	•	–	27.49	95.51	75.67
4	–	–	•	30.89	96.26	78.89
5	•	•	–	31.10	96.40	79.06
6	•	–	•	34.57	96.95	81.64
7	–	•	•	28.94	96.26	77.93
8	•	•	•	33.00	96.84	80.82

corelează cu performanța sa de estimare a încrederii. Printre cele trei metode de îmbunătățire propuse, observăm că scalarea temperaturii oferă cea mai mare creștere a performanței pentru toate cele trei valori (rândul 2). În mod surprinzător, metoda dropout-ului îmbunătățește numai performanța AUPR față de linia de bază (rândul 3). În combinațiile de două metode, scalarea temperaturii și ansamblul se completează reciproc și obțin performanțe mai bune.

În continuare, discutăm pe scurt diferite perspective legate de metodologia propusă.

Mărirea setului de caracteristici. Caracteristicile pe care le-am investigat au avantajul de a fi generale, pentru că valorifică probabilitățile la nivel de token, care sunt disponibile în majoritatea, dacă nu în toate, utilitățile pentru RAV-uri de tip end-to-end existente. Cu toate acestea, setul de caracteristici ar putea fi extins cu probabilități pentru sunetul de intrare sau textul generat, sau cu informații despre durata cuvântului, care pot fi extrase din ponderile de atenție.

Învățarea caracteristicilor. Inspirați de studiile precedente [Corbière, 2019; Chen, 2019], am experimentat, de asemenea, cu învățarea unei rețele de estimare a încrederii pe baza caracteristicilor extrase din modelul end-to-end (mai exact, activări logit și pre-logit). Cu toate acestea, experimentele noastre nu au reușit să arate îmbunătățiri față de rezultatele prezentate.

Tratarea cuvintelor pierdute. Pentru a genera etichete pentru scorurile de încredere, aliniem textul de referință la textul prezis și marcăm cuvintele corecte din textul prezis ca pozitive și substituțiile și inserțiile ca negative. Această abordare este tipică în literatura de evaluare a scorurilor de încredere, dar ratează erorile făcute prin ștergerea (pierderea) cuvintelor din referință. Mai multe lucrări au abordat această problemă [Seigel, 2014; Ragni, 2018]; lăsăm pentru viitor să extindem abordarea curentă pentru această sarcină.

6. Setup-ul experimental pentru experimente pe limba română

Pentru experimentele pe limba română folosim seturile de date de evaluare prezentate în secțiunea 2.1.1 a raportului proiectului component TADARAV [Georgescu, 2020]: RSC-eval (5.5 ore de vorbire citită) și SSC-eval1+2 (3.5+1.5 ore de vorbire spontană).

Sistemul de transcriere de vorbire folosit pentru aceste experimente este cel introdus în secțiunea 2.2.1 a raportului proiectului component TADARAV [Georgescu, 2020]. Sistemul RAV este implementat folosind librăria ESPnet și se bazează pe o arhitectură de tip Transformer. Vocabularul sistemului cuprinde 1,000 de subcuvinte și a fost extras din transcrierile materialelor audio cu care a fost antrenat întregul sistem: cele 517 ore de vorbire din seturile de date RSC-train, SSC-train1+2 și SSC-train3+4-trans-v4. Pentru decodare se folosește un model de limbă adițional implementat tot ca Transformer și antrenat pe cele ~352M de cuvinte din seturile de date news2017 și talkshows. Performanța sistemului de RAV folosit, exprimată sub forma erorii la nivel de cuvânt (WER), este de 3.4% WER pe RSC-eval, 15.3% WER pe SSC-eval1, respectiv 23.5% WER pe SSC-eval2.

Din cauza complexității modelului CTC-Transformer și timpului mare de inferență pe configurația hardware disponibilă în acest proiect, nu s-au mai putut antrena modele adiționale cu inițializari diferite ale ponderilor, pentru ansamblurile de modele.

Metricile de evaluare introduse în secțiunea anterioară (*i.e.* AUROC, AUPRs, AUPRs) sunt universale, independente de limbă și, în consecință, vor fi folosite și pentru experimentele pe limba română.

7. Rezultate experimentale pentru limba română

Pentru limba română am reluat următoarele experimente realizate și pentru sistemul de estimare a încrederii pentru limba engleză:

- evaluarea diverselor caracteristici ce pot fi utilizate pentru estimarea încrederii (Tabelul 5);
- evaluarea diverselor metode de agregare a caracteristicilor la nivel de subcuvânt pentru a obține scoruri de încredere la nivel de cuvânt (Tabelul 5);
- evaluarea tehnicii dropout pentru îmbunătățirea scorurilor de încredere de mai sus (Tabelul 6).

7.1 Caracteristici și metode de agregare

În această secțiune evaluăm caracteristicile pentru reprezentarea incertitudinii din subcuvinte și tehnicile de agregare propuse pe cele trei seturi de date prezentate în secțiunea 2.1.1 a raportului proiectului component TADARAV [Georgescu, 2020]: RSC-eval (5.5 ore de vorbire citită) și SSC-eval1+2 (3.5+1.5 ore de vorbire spontană). Tabelul 5 prezintă rezultatele pentru toate combinațiile de caracteristici și agregări.

Comparație a caracteristicilor. În ceea ce privește tipul de caracteristici (log-proba vs. neg-entropy), rezultatele pe seturile de date în limba română indică exact aceleași concluzii ca și în cazul limbii engleze:

- pe seturile de date pe care sistemul RAV are o acuratețe relativ bună (în cazul nostru RSC-eval și SSC-eval1) caracteristicile de tip probabilități log obțin rezultate mai bune decât caracteristicile pe bază de entropie indiferent de agregare
- pe seturile de date mai dificile (SSC-eval2 pentru română, respectiv CommonVoice pentru engleză) caracteristicile bazate pe entropie agregate cu metoda *sum* prezintă rezultate similare cu, uneori chiar mai bune decât caracteristicile de tip probabilități log.

Comparație a metodelor de agregare. În ceea ce privește metodele de agregare, concluziile sunt relativ diferite pentru limba română:

- pe limba română agregarea minimul funcționează mai bine indiferent de caracteristici, în timp ce pe limba engleză caracteristicile de tip probabilități log erau agregate mai bine folosind suma;
- agregarea prin medie este, în general, slabă pentru ambele caracteristici, sugerând că măsurile invariante în lungime sunt dăunătoare;

Tabelul 5. Rezultate pentru evaluarea scorurilor de încredere pentru toate combinațiile de caracteristici (*caract.*) și metode de agregare (*agg.*), pe cele trei seturi de evaluare în limba română. Pentru toate cele trei metrici de evaluare utilizate (AU-Pre, AU-PRs, AU-ROC) valorile mai mari indică performanță mai bună.

		RSC-eval / 1.8%			SSC-eval1 / 11.0%			SSC-eval2 / 14.0%		
caract.	agg.	AU-Pre	AU-PRs	AU-ROC	AU-Pre	AU-PRs	AU-ROC	AU-Pre	AU-PRs	AU-ROC
log-proba	sum	15.81	98.77	74.50	27.19	96.14	74.57	15.79	91.90	61.52
	min	18.30	98.91	78.66	28.32	96.34	77.24	13.86	91.99	60.46
	avg	16.22	98.74	75.75	26.69	96.04	75.38	14.04	92.09	60.60
neg-entropy	sum	15.82	98.73	75.65	26.44	96.03	75.31	14.30	92.17	61.09
	min	16.28	98.86	77.79	27.10	96.13	76.14	12.75	91.47	57.94
	avg	13.55	98.65	74.02	24.97	95.75	73.62	13.37	91.68	58.88

- excepții de la concluziile de mai sus apar din nou pentru seturile de date mai dificile (SSC-eval2 pentru română, respectiv CommonVoice pentru engleză) pentru care agregarea prin sumă produce cele mai bune rezultate, indiferent de caracteristicile utilizate.

Comparație între seturile de date. Așa cum era de așteptat, modelul preantrenat are cea mai bună performanță în ceea ce privește datele din domeniu (1.8% WER pe RSC-eval și 11.0% pe SSC-eval1) și performanță mai slabă pe date din afara domeniului (14.0% pe SSC-eval2). În fiecare dintre aceste setări, numărul de cuvinte care sunt clasificate corect se schimbă, trecând de la mai multe pe setul RSC la mai puține pe seturile de date SSC. Această observație explică de ce performanța pentru AUPRs este mare pe RSC și scade pe SSC-eval1 și 2, și, dimpotrivă, de ce performanța pentru AUPRe este mai slabă pentru RSC și mai mică pentru SSC-eval1. Din păcate, din exact acest motiv—performanța diferită a sistemului RAV de bază pe cele patru seturi de date—este imposibil de comparat metodele de încredere între seturile de date. Performanța pentru AUPRe pe setul de date SSC-eval2 face din nou notă discordantă: dat fiind faptul că sistemul RAV are performanțe slabe de transcriere pe acest set de date ne-am fi așteptat ca acele cuvinte transcrise greșit să poată fi foarte ușor de identificat (un AUPRe de valoare mare), însă acest lucru nu se întâmplă.

7.2 Tehnica de dropout

În această secțiune evaluăm metodele de estimare a scorurilor de încredere după îmbunătățirea probabilităților subcuvintelor prin tehnica de dropout. Această tehnică presupune transcrierea datelor de intrare de N=4, 8, 16, 32, respectiv 64 de ori folosind dropout. Mediem apoi cele N predicții independente. Tabelul 6 prezintă rezultatele pentru toate combinațiile de caracteristici și agregări și dropout cu N=64.

Comparând rezultatele din Tabelul 5 cu rezultatele din Tabelul 6 observăm îmbunătățiri pentru seturile de date RSC-eval și SSC-eval1, însă o scădere de performanță pentru setul de date SSC-eval2. Nici îmbunătățirile obținute pe RSC-eval și SSC-eval1 nu sunt foarte mari (AU-ROC crește de la 78.66 la 82.02 pe RSC-eval și de la 77.24 la 79.90 pe SSC-eval1).

Toate metricile de performanță se îmbunătățesc pe măsură ce sunt mediate mai multe rezultate de transcriere (pe măsură ce N crește) și, la fel ca și pentru limba engleză, creșterile de performanță se plafonează pentru N=64.

Un ultim considerent demn de menționat este acela că pentru a utiliza tehnica dropout în estimarea încrederii este nevoie de un timp de calcul net superior situației inițiale. Practic, pentru N=64 trebuie realizate 64 de transcrieri, deci timpul de calcul va fi de aproximativ 64 de ori mai mare. În concluzie, avantajele minime obținute prin utilizarea acestei tehnici pentru îmbunătățirea probabilităților subcuvintelor sunt contrabalansate de un dezavantaj semnificativ ce ține de timpul de calcul.

Tabelul 6. Rezultate pentru evaluarea tehnicii dropout (N=64) pentru îmbunătățirea scorurilor de încredere.

caract.	agg.	RSC-eval / 1.8%			SSC-eval1 / 11.0%			SSC-eval2 / 14.0%		
		AU-Pre	AU-PRs	AU-ROC	AU-Pre	AU-PRs	AU-ROC	AU-Pre	AU-PRs	AU-ROC
log-proba	sum	17.76	98.97	77.14	28.96	96.85	77.31	16.03	90.08	58.47
	min	18.90	99.17	82.02	29.01	97.10	79.90	13.58	90.73	56.77
	avg	15.73	99.13	81.02	27.75	96.95	79.22	13.02	90.78	56.22
neg-entropy	sum	9.53	98.75	71.19	20.40	96.20	71.68	15.18	90.45	56.80
	min	14.43	99.11	80.63	27.01	96.90	78.86	11.85	90.22	53.74
	avg	11.83	99.02	78.71	25.50	96.70	77.84	11.94	90.36	53.85

8. Utilizarea scorurilor de încredere propuse pentru generarea de date adnotate

În activitatea 2.12 din etapa 2019 am propus utilizarea scorurilor de încredere la nivel de cuvânt puse la dispoziție de un sistem de RAV pentru a selecta cuvintele presupus transcrise corect de acesta. Selecția se face impunând un prag minim pentru scorul de încredere la nivel de cuvânt, prag sub care cuvântul în cauză este considerat ca fiind transcris greșit. Secvențele de cuvinte presupus transcrise corect, împreună cu datele audio corespunzătoare, ar urma să formeze un nou set de date adnotat. Pentru ca aplicarea acestei metode să genereze seturi de date adnotate suficient de mari, este necesar ca setul de date neadnotat utilizat la intrarea sistemului să aibă o dimensiune considerabilă. De exemplu, în acest proiect am utilizat seturile de date SSC-train3-raw și SSC-train4-raw care însumează 913 ore de vorbire. Metoda descrisă mai sus se poate aplica numai în cazul în care transcrierea acestui set de date neadnotat cu ajutorul sistemului de RAV se poate realiza în mod eficient. Din păcate experimentele noastre au arătat că sistemul RAV ESPnet transcrie o oră de vorbire într-un interval de 2 până la 7 ore, în funcție de cât de similară este acustica respectivei ore de vorbire raportat la setul de date de antrenare. În acest context, transcrierea setului de date de 913 ore ar fi durat între 76 de zile și 266 de zile. În consecință, din considerente ce țin de infrastructura de calcul, nu am putut utiliza noile metode de estimare a scorurilor de încredere pentru generare de date audio adnotate.

Bibliografie

- [Ardila, 2020] Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber, “Common Voice: A massively-multilingual speech corpus,” in International Conference on Language Resources and Evaluation, 2020.
- [Ashukha, 2020] Arsenii Ashukha, Alexander Lyzhov, Dmitry Molchanov, and Dmitry Vetrov, “Pitfalls of in-domain uncertainty estimation and ensembling in deep learning,” in International Conference on Learning Representations, 2020.
- [Chen, 2019] Tongfei Chen, Jirí Navrátil, Vijay Iyengar, and Karthikeyan Shanmugam, “Confidence scoring using whitebox meta-models with linear classifier probes,” in International Conference on Artificial Intelligence and Statistics, 2019, pp. 1467–1475.
- [Corbière, 2019] Charles Corbière, Nicolas Thome, Avner Bar-Hen, Matthieu Cord, and Patrick Pérez, “Addressing failure prediction by learning model confidence,” in Advances in Neural Information Processing Systems, 2019, pp. 2902–2913.
- [Cortina, 2016] Isaías Sánchez Cortina, Jesús Andrés-Ferrer, Alberto Sanchis, and Alfons Juan, “Speaker-adapted confidence measures for speech recognition of video lectures,” *Computer Speech & Language*, vol. 37, pp. 11–23, 2016.
- [Del-Agua, 2018] M. A. Del-Agua, A. Gimenez, A. Sanchis, J. Civera, and A. Juan, “Speaker-adapted confidence measures for ASR using deep bidirectional recurrent neural networks,” *Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 7, pp. 1198–1206, 2018.
- [Errattahi, 2018] Rahhal Errattahi, Salil Deena, Asmaa El Hannani, Hassan Ouahmane, and Thomas Hain, “Improving ASR error detection with RNNLM adaptation,” in IEEE Spoken Language Technology Workshop, 2018, pp. 190–196.
- [Gal, 2016] Yarin Gal and Zoubin Ghahramani, “Dropout as a Bayesian approximation: Representing model uncertainty in deep learning,” in International Conference on Machine Learning, 2016, pp. 1050–1059.
- [Georgescu, 2020] A.-L. Georgescu, A. Caranica, C. Manolache, G. Pop, D. Oneață, H. Cucu, C. Burileanu, D. Burileanu, „*Proiect component TADARAV: Raport științific și tehnic în extenso 2020*”.
- [Guo, 2017] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger, “On calibration of modern neural networks,” in International Conference on Machine Learning, 2017, pp. 1321–1330.
- [Hadian, 2018] Hossein Hadian, Hossein Sameti, Daniel Povey, and Sanjeev Khudanpur, “End-to-end speech recognition using lattice-free MMI,” in Interspeech, 2018, pp. 12–16.
- [Hazen, 2002] Timothy J Hazen, Stephanie Seneff, and Joseph Polifroni, “Recognition confidence scoring and its use in speech understanding systems,” *Computer Speech & Language*, vol. 16, no. 1, pp. 49–67, 2002.

- [Hendrycks, 2016] Dan Hendrycks and Kevin Gimpel, “A baseline for detecting misclassified and out-of-distribution examples in neural networks,” in International Conference on Learning Representations, 2016.
- [Hinton, 2015] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean, “Distilling the knowledge in a neural network,” arXiv preprint arXiv:1503.02531, 2015.
- [Kalgaonkar, 2015] Kaustubh Kalgaonkar, Chaojun Liu, Yifan Gong, and Kaisheng Yao, “Estimating confidence scores on ASR results using recurrent neural networks,” in IEEE International Conference on Acoustics, Speech and Signal Processing, 2015, pp. 4999–5003.
- [Karita, 2019] Shigeki Karita, Nanxin Chen, Tomoki Hayashi, Takaaki Hori, Hirofumi Inaguma, Ziyang Jiang, Masao Someki, Nelson Enrique Yalta Soplín, Ryuichi Yamamoto, Xiaofei Wang, Shinji Watanabe, Takenori Yoshimura, and Wangyou Zhang, “A comparative study on transformer vs RNN in speech applications,” in Workshop on Automatic Speech Recognition and Understanding, 2019, pp. 449–456.
- [Kemp, 1997] Thomas Kemp and Thomas Schaaf, “Estimating confidence using word lattices,” in Eurospeech, 1997.
- [Kudo, 2018] Taku Kudo, “Subword regularization: Improving neural network translation models with multiple subword candidates,” in Association for Computational Linguistics, 2018, pp. 66–75.
- [Lakshminarayanan, 2017] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell, “Simple and scalable predictive uncertainty estimation using deep ensembles,” in Advances in Neural Information Processing Systems, 2017, pp. 6402–6413.
- [Li, 2019] Qiuqiang Li, PM Ness, Anton Ragni, and Mark JF Gales, “Bi-directional lattice recurrent neural networks for confidence estimation,” in IEEE International Conference on Acoustics, Speech and Signal Processing, 2019, pp. 6755–6759.
- [Lüscher, 2019] Christoph Lüscher, Eugen Beck, Kazuki Irie, Markus Kitzka, Wilfried Michel, Albert Zeyer, Ralf Schlüter, and Hermann Ney, “RWTH ASR systems for LibriSpeech: Hybrid vs attention,” in Interspeech, 2019, pp. 231–235.
- [Malinin, 2020] Andrey Malinin and Mark Gales, “Uncertainty in structured prediction,” arXiv preprint arXiv:2002.07650, 2020.
- [Ogawa, 2017] Atsunori Ogawa and Takaaki Hori, “Error detection and accuracy estimation in automatic speech recognition using deep bidirectional recurrent neural networks,” *Speech Communication*, vol. 89, pp. 70–83, 2017.
- [Ovadia, 2019] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek, “Can you trust your model’s uncertainty? Evaluating predictive uncertainty under dataset shift,” in Advances in Neural Information Processing Systems, 2019, pp. 13991–14002.
- [Panayotov, 2015] V. Panayotov, G. Chen, D. Povey and S. Khudanpur, “*Librispeech: An ASR corpus based on public domain audio books*,” in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5206–5210, 2015.
- [Park, 2019] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le, “SpecAugment: A simple data augmentation method for automatic speech recognition,” in Interspeech, 2019, pp. 2613–2617.
- [Povey, 2011] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely, “The Kaldi speech recognition toolkit,” in Workshop on Automatic Speech Recognition and Understanding, 2011.
- [Ragni, 2018] Anton Ragni, Qiuqiang Li, Mark JF Gales, and Yongqiang Wang, “Confidence estimation and deletion prediction using bidirectional recurrent neural networks,” in IEEE Spoken Language Technology Workshop, 2018, pp. 204–211.
- [Rousseau, 2014] Anthony Rousseau, Paul Deléglise, and Yannick Estève, “Enhancing the TED-LIUM corpus with selected data for language modeling and more TED talks,” in International Conference on Language Resources and Evaluation, 2014, pp. 3935–3939.

- [Seigel, 2013] Mathew Stephen Seigel, Confidence estimation for automatic speech recognition hypotheses, Ph.D. thesis, University of Cambridge, 2013.
- [Seigel, 2014] Matthew Stephen Seigel and Philip C Woodland, "Detecting deletions in ASR output," in IEEE International Conference on Acoustics, Speech and Signal Processing, 2014, pp. 2302–2306.
- [Sperber, 2017] Matthias Sperber, Graham Neubig, Jan Niehues, and Alex Waibel, "Neural lattice-to-sequence models for uncertain inputs," in Empirical Methods in Natural Language Processing, 2017, pp. 1380–1389.
- [Srivastava, 2014] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 1 2014.
- [Swarup, 2019] Prakhar Swarup, Roland Maas, Sri Garimella, Sri Harish Mallidi, and Björn Hoffmeister, "Improving ASR confidence scores for Alexa using acoustic and hypothesis embeddings," in Interspeech, 2019, pp. 2175–2179.
- [Tüske, 2019] Zoltán Tüske, Kartik Audhkhasi, and George Saon, "Advancing sequence-to-sequence based speech recognition," in Interspeech, 2019, pp. 3780–3784.
- [Vaswani, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.
- [Vesely, 2013] Karel Vesely, Mirko Hannemann, and Lukas Burget, "Semi-supervised training of deep neural networks," in *Workshop on Automatic Speech Recognition and Understanding*, 2013, pp. 267–272.
- [Vyas, 2019] Apoorv Vyas, Pranay Dighe, Sibio Tong, and Hervé Bourlard, "Analyzing uncertainties in speech recognition using dropout," in IEEE International Conference on Acoustics, Speech and Signal Processing, 2019, pp. 6730–6734.
- [Watanabe, 2018] Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplín, Jahn Heymann, Matthew Wiesner, Nanxin Chen, Adithya Renduchintala, and Tsubasa Ochiai, "ESPnet: End-to-End Speech Processing Toolkit," in *Proc. Interspeech*, pp. 2207-2211, 2018.
- [Weintraub, 1997] Mitch Weintraub, Françoise Beaufays, Ze'ev Rivlin, Yochai König, and Andreas Stolcke, "Neural-network based measures of confidence for word recognition," in IEEE International Conference on Acoustics, Speech and Signal Processing, 1997, vol. 2, pp. 887–890.
- [Yu, 2010] Dong Yu, Balakrishnan Varadarajan, Li Deng, and Alex Acero, "Active learning and semi-supervised learning for speech recognition: A unified framework using the global entropy reduction maximization criterion," *Computer Speech & Language*, vol. 24, no. 3, pp. 433–444, 2010.