Studiu privind starea artei: Estimarea scorurilor de încredere pentru sistemele de recunoaștere automată a vorbirii

Ana Neacșu Dan Oneață Horia Cucu
Corneliu Burileanu Dragoș Burileanu

Cuprins

1 Introducere

Recunoașterea automată a vorbirii (RAV) ne permite să interacționăm cu sistemele din jurul nostru într-un mod facil și natural, prin intermediul comenzilor vocale. În ultimii ani, s-au dezvoltat o multitudine de aplicații interactive construite pe această tehnologie, fiind una dintre cele mai comode interfețe om-mașină. Exemple de astfel de aplicații includ cele care accesează și filtrează conținut de pe Internet (ex. Siri, Cortana), controlul unei interfețe, conversia automată a textului în scris, etc. În afară de aceste aplicații specifice utilizatorului, serviciile care efectuează transcrierea pe scară largă a conținutului audio ce conține vorbire și cele care semnalează apariția anumitor cuvinte cheie, se bazează, de asemenea, pe recunoaștere automată a vorbirii. Interesul ridicat în cercetarea acestei tehnologii a dus la îmbunătățirea constantă a performanțelor de-a lungul timpului, în măsura în care sistemele state-of-the-art, bazate pe modele statistice sofisticate de vorbire și de limbă ating niveluri impresionante de precizie.

Cu toate acestea, sistemele de ultimă generație nu sunt încă perfecte și, adesea, produc transcrieri care conțin erori de recunoaștere. Aceste erori sunt produse, în principal, de doi factori. În primul rând, mediul acustic în care au fost capturate datele care sunt folosite pentru estimarea parametrilor modelului acustic poate fi foarte diferit de cel folosit de utilizator în mediul real. În general, există diferențe mari intre mediul de antrenament și cel de testare, cum ar fi condiții de zgomot diferite, vorbitori diferiți, alt tip de canal, etc. În al doilea rând, ipotezele care se fac în mod tipic în modelele statistice de bază, astfel încât aplicarea lor să fie adaptivă, sunt adesea prea restrictive.

Având în vedere că există un anumit grad de incertitudine în exactitatea transcrierilor rezultate dintr-un sistem RAV, este important să avem o modalitate de a măsura câtă încredere putem avea că ieșirea sistemului este corectă. Gradul de încredere în transcrierea unui sistem RAV poate fi măsurat cu ajutorul scorurilor de încredere. Acestea se pot folosi în multe sarcini ce presupun recunoașterea vorbirii sau a vorbitorului. De exemplu, sistemele de dialog se dovedesc a fi mai capabile de a răspunde la solicitarea utilizatorului folosind aceste măsuri și pot ghida dialogul în consecință. În sistemele de transcriere semi-automată, segmentele audio cu nivel scăzut de încredere pot fi transcrise manual sau aruncate în timpul procedeului de selecție de date, astfel având siguranța că păstrăm doar traducerile corecte.

Procesul de obținere a măsurilor de încredere pentru transcrierile ipotetice

de la ieșirea RAV este cunoscută sub numele de Estimare a Încrederii (EI). Tehnicile de EI sunt îmbunătățite continuu astfel încât să producă măsuri concludente mai precise și mai fiabile care pot fi utilizate pentru a îmbunătăți performanța aplicațiilor care utilizează tehnologia RAV.

Literatura de specialitate existentă în acest domeniu împarte EI în trei clase de abordări Jiang (2005). În prima, măsura de încredere pentru un cuvânt este definită ca fiind probabilitatea aposteriori a cuvântului transcris. Aceste probabilități pot fi obținute în mai multe moduri: 1-Best decoding Evermann and Woodland (2000), Wessel et al. (2001), consensus network (CN) clustering Mangu et al. (2000), minimum Bayes risk (MBR) decoding Goel and Byrne (2000), Xu et al. (2010). Totuși, probabilitățile rezultate, de cele mai multe ori, nu sunt măsuri precise ale gradului de încredere, ci tind să supraestimeze probabilitatea ca respectivul cuvânt să fie corect. Prin urmare, pentru a obține rezultate relevante este necesară o formă de mapare a scorurilor de încredere obținute. Cea de-a doua abordare formulează sarcina de EI ca o ipoteză statistică de testare, raportul dintre dintre probabilitatea ca transcrierea să se fi realizat corect și probabilitatea ca transcrierea să fie eronată. Asta implică antrenarea mai multor modele, ceea ce reprezintă o creștere semnificativă a efortului de proiectare și a complexității sistemului. Cea de-a treia abordare se concentrează asupra implementării unui clasificator pentru a estima probabilitatea unei transcrieri corecte, pe baza unui set de trăsături predictive (cu rol informativ) ce procesează caracteristici fundamentale ale sistemului RAV. Această abordare este mai flexibilă, întrucât permite combinarea informațiilor din surse diferite, îmbunătățind procesul de estimare a încrederii.

2 Metode de estimare a încrederii

Metodele de EI pot fi împărțite în trei mari categorii, ce vor fi tratate separat, urmând ca în final să fie comparate în secțiunea de concluzii. Deoarece majoritatea literaturii existente se referă la estimarea încrederii la nivel de cuvânt, aceasta va fi forma generală asumată în secțiunile următoare.

2.1 Abordarea bazată pe Clasificare

Tehnicile de EI care se încadrează în această categorie se bazează pe implementarea unui clasificator la ieșirea sistemului RAV. Scopul principal al

acestui clasificator este de a estima probabilitatea ca un cuvânt să fi fost transcris corect. Principiul abordării implică clasificarea cuvintelor de la ieșirea sistemului pe baza unor trăsături, care pot și extrase din numeroase surse. Una dintre aceste surse este sistemul RAV propriu-zis. Caracteristicile rezultate încorporează informații referitoare la procesul de recunoaștere în momentul în care apare un cuvânt ipotetic.

2.1.1 Trăsături predictive

Trăsăturile utilizate în clasificarea unor cuvinte ipotetice ca fiind corecte sau incorecte se numesc trăsături predictive. Multe astfel de trăsături au fost propuse în literatură, dintre care unele sunt derivate din informațiile care sunt reprezentate de laticele de recunoaștere. Laticele de recunoaștere constau în reprezentări compacte a celor mai relevante ipoteze generate în timpul recunoașterii. De obicei, arcele reprezintă cuvinte, iar nodurile reprezintă probabilitatea de apariție a fiecărui cuvânt. Teoretic, caracteristica unui predictor optim este o măsură pentru care distribuția cuvintelor recunoscute corect diferă semnificativ de cea a cuvintelor incorecte. În continuare sunt descrise cele mai cunoscute tipuri de trăsături predictive.

- Trăsături bazate pe structura rețelei: Densitatea ipotezei Kemp and Schaaf (1997) este o măsură bazată pe presupunerea ca numărul de arce alternative ce se întind pe segmentul de timp pentru un cuvânt în cea mai probabilă traducere este un indicator al incertitudinii RAV-ului în acea ipoteză.

- Trăsături bazate pe informația acustică: Scorurile de probabilitate acustică, normalizate la numărul de ferestre sau la numărul de foneme Pinto and Sitaram (2005), este o reprezentare a legăturii dintre semnalul acustic și cuvântul ipotetic. Stabilitatea acustică este o măsură dată de numărul aparițiilor unui cuvânt dat pe aceeași poziție în K ieșiri diferite ale sistemului RAV. Fiecare dintre cele K ieșiri diferite este scalată cu un factor GSF (Grammar Scaling Factor), notat cu γ. Acesta are ca efect principal diminuarea co-dependenței dintre modelul de limbă și modelul acustic. Principiul acestui tip de trăsături se bazează pe observația că acele cuvinte ipotetice care sunt aliniate pe aceeași poziție pentru valori diferite ale factorului γ au probabilitatea mai mare să fie în concordanță cu mesajul vorbit. În situațiile în care se utilizează un model discriminativ, se pot extrage trăsături din

probabilitățile a posteriori calculate la nivel de fereastră sau la nivel sub-cuvânt.

- Trăsături bazate pe modelul de limbă: Willett et al. (1998) a arătat că performața EI se îmbunătățește prin combinarea trăsăturilor pur acustice cu trăsături extrase pe baza modelului de limbă prin acordarea unor probabilități de apariție unor cuvinte sau secvențe de cuvinte. Au fost propuse și măsuri non-probabilistice bazate pe modelul de limbă, cum ar fi soluția propusă în Weintraub et al. (1997).

- Trăsături bazate pe durata cuvântului: Caracteristicile extrase din numărul de foneme ce compun cuvântul ipotetic pot fi estimate din laticea de recunoștere. Logica din spatele acestui tip de trăsături rezidă în observația că de obicei secvențele de cuvinte mai scurte corespund unor regiuni din semnal pe care modelul acustic are performanțe mai slabe.

- Trăsături la nivelul secvenței de cuvinte rostite: Hazen et al. (2002) a propus un set de trăsături ce pot fi extrase la nivel de cuvânt sau secvență de cuvinte rostite. Aceste trăsături includ poziția cuvântului, lungimea secvenței, identitatea lexicală a cuvântului ipotetic etc.

2.1.2 Clasificatori folosiți pentru generarea scorurilor de încredere

Majoritatea trăsăturilor predictive caracterizează semnalul vocal în mod asemănător, acestea oferind utilizatorului o metrică pe baza căreia se poate lua decizia dacă un cuvânt a fost transcris corect sau nu. Totuși, aceste trăsături nu sunt identice, fiecare analizând semnalul din altă perspectivă. Din acest motiv, cea mai bună abordare este aceea de a combina diferite trăsături predictive, mărind astfel puterea discriminativă a fiecărei trăsături în parte și obținând un sistem de estimare a încrederii mai performant. Mare parte a literaturii care abordează acest tip de EI se concentrează pe găsirea unei combinații optime de trăsături. O lucrare reprezentativă care explorează definirea, combinarea și evaluarea unor trăsături predictive pentru Estimarea Încrederii este Chase (1997). Câțiva clasificatori care au folosiți pentru a găsi combinații optime de trăsături predictive includ: arbori de decizie Neti et al. (1997), rețele neurale Weintraub et al. (1997), vectori suport Zhang and Rudnicky (2001) etc.

Dacă cercetarea din anii '90-'00 s-a bazat pe găsirea unor trăsături cât mai relevante, în ultimii ani, odată cu evoluția sistemelor de calcul, cerceta-

rea din acest domeniu s-a concentrat mai degrabă pe dezvoltarea unor clasificatori cât mai performanți, decât pe găsirea altor trăsături. Seigel (2013) adresează problema clasificării folosind CRF (*Conditional Random Fields*), un model statistic ce calculează probabilitatea ca ieșirea RAV-ului să fie corectă, considerând secvența de intrare. Această metodă a adus îmbunătățiri semnificative în domeniul estimării încrederii datorită faptului că sistemul propus exploatează natura secvențială a vorbirii. Alte lucrări ce abordează clasificarea folosind CRF sunt: Fayolle et al. (2010), Yu et al. (2010), Marin et al. (2012)

Rețelele neurale recurente (*RNNs- Recurrent Neural Networks*) au fost recent introduse cu succes în domeniul prelucrării semnalului vocal, mai cu seamă în sistemele de recunoaștere automată a vorbirii. Astfel de arhitecturi au fost propuse de Mikolov et al. (2010), Mikolov et al. (2011), Hori et al. (2014). Kalgaonkar et al. (2015) evaluează modele de tip RNN pe trăsături extrase din latice (scor acustic, scor model de limbă, durată, scor zgomot, perplexitatea a modelului de limbă, etc.). Contrastează modelul de tip RNN cu un MLP: obțin îmbunatățiri relative de aproximativ 10% pentru TPR. Un rezultat interesant e că pentru primul cuvânt dintr-o rostire rezultatele cu RNN sunt ceva mai slabe decât cu MLP; autorii motivează asta pe baza faptului că RNN-ul nu s-a putut folosi încă de context. Afirmația este sprijinită de observația că începând cu cel de-al șaselea cuvânt se observă îmbunătățiri.

În Ogawa and Hori (2017) este abordată o arhitectură de tip RNN bidirecțional BiDRNN (*deep bidirectional RNNs*; DBRNNs) pentru detecția erorilor în sitemele RAV. Sistemul BiDRNN depășește performanțele atinse de cele bazate pe probabilități posteriori, CRF, alte structuri de rețele neurale și pe RNN-urile unidirecționale (UniDNN). CRF-urile introduc o creștere semnificativă a performanțelor față de sistemele bazate pe probabilități posteriori, având raportat un indice NCE (*Normalized Cross Entropy*) de 0.363, față de 0.130 (pentru posteriori). Utilizarea rețelelor neurale profunde (DNN) sau UniDRNN-urilor nu aduce o îmbunătățire semnificativă, având un NCE raportat de 0.366, respectiv 0.382. Utilizarea BiDRNN-urilor aduce totuși o creștere notabilă, cu un NCE de 0.404. Rezultatele obținute sunt o consecință a faptului că rețelele recurente bidirecționale pot lua în considerare contextul bidirecțional mai lung și pot modela relații puternic neliniare între vectorii de intrare și etichetele de ieșire. Printre trăsăturile utilizate în procesul de clasificare se numără: numărul de ferestre, numărul de foneme, partea de vorbire, numărul de ipoteze alternative, logaritmul probabilității unigram

și trigram, etc.

Del-Agua et al. (2018) prezintă, în mod similar cu Ogawa and Hori (2017), o metodă de clasificare bazată pe BiDRNN-uri. Deosebirea constă în faptul că adaptează un RNN general la vorbitor prin tehnica de fine-tuning: se continuă procesul de antrenare, dar cu o rată de învățare mică. Și în această lucrare este confirmată superioritatea BiDRNN-urilor față de CRF și metode bazate pe probabilitatea aposteriori. Rezultatele obținute sunt asemănătoare cu cele raportate de Ogawa and Hori (2017), diferența relativă dintre indicele NCE înregistrat utilizând sisteme CRF și cel obținut folosind BiDRNN-uri fiind aceeași. Totuși, procesul de adaptare nu aduce îmbunătățiri semnificative sistemului.

2.1.3 Dezavantaje

S-a arătat că prin combinarea mai multor trăsături predictive se poate îmbunătăți performanța sistemelor ce măsoară încrederea. Această îmbunătățire este semnificativă atunci când trăsăturile sunt statistic independente între ele. Un studiu în Kemp and Schaaf (1997) a arătat că majoritatea trăsăturilor predictive sunt puternic corelate între ele. Din acest motiv, încercările de a combina trăsăturile nu schimbă semnificativ performanțele sistemului, față de cazul în care pentru proiectarea sistemului se consideră doar cea mai relevantă trăsătură.

2.2 Abordarea bazată pe probabilitatea posteriori

Sistemele moderne de RAV au ca scop recunoașterea corectă a vorbirii folosind o abordare în care se aplică tehnici statistice de recunoaștere a formelor. Dacă sistemele de recunoaștere ar fi perfecte, măsurarea încrederii nu ar mai fi necesară. Scalarea și presupunerile inițiale făcute într-un sistem de RAV sunt cauzele principale din spatele traducerilor eronate. Totuși, tot datorită naturii statistice a sistemelor de recunoaștere, se poate considera că alegerea unui cuvânt ipotetic poate oferi o indicație asupra acurateței transcrierii. Acest punct de vedere formează baza abordării a posteriori descrisă în această secțiune.

Regula de decizie în sistemele RAV este de a alege cuvântul ipotetic ca fiind cel cu probabilitatea a posteriori cea mai mare. Datorită constrângerilor impuse de sistemele bazate pe HMM, probabilitatea posteriori trebuie estimată, fie prin asumarea unor presupuneri inițiale în legătură cu forma

distribuției $p(X)$, fie prin utilizarea unor metode de aproximare a distribuției.

În prima abordare, așa numitele metode bazate pe umplere *(en.. filler based- methods)* se bazează pe utilizarea unui set de modele de umplere simple și cu constrângeri fixe pentru a reprezenta distribuția dorită, plecând de la premisa că avem o idee despre cum ar trebui să arate. Exemple de astfel de sisteme includ metode ce utilizează modele de recunoaștere a fonemelor Young and Woodland (1994), modele *catch all* Kamppari and Hazen (2000) etc. Totuși, aceste tehnici impun niște presupuneri destul de largi și de aceea aplicarea lor nu aduce îmbunătățiri substanțiale ale performanței sistemului.

Cea de-a doua abordare se bazează pe metode de aproximare pentru a estima adevărata distribuție $P(X)$. Această abordare s-a dovedit a fi mult mai performantă. Un estimat bun al distribuției se poate obține calculând probabilitatea posteriori pe laticea de recunoaștere. Deși oferă rezultate mai bune, este o sarcină foarte costisitoare din punct de vedere computațional. Pentru simplificare, în sistemele implementate în Rueber (1997), Wessel et al. (1998) au fost luate în considerare doar cele mai bune N ipoteze.

2.2.1 Probabilități posteriori bazate pe timp

O tehnică pentru estimarea probabilității posteriori la nivel de cuvânt bazată pe forma latice a posibilelor ipoteze a fost propusă de (Evermann and Wooland, 2000)). Prima etapă în acest algoritm este calcularea probabilității posteriori a fiecărui arc în latice. Asemănările găsite de modelul lingvistic (LM) și de cel acustic (AM) sunt stocate în latice și utilizate pentru a calcula distribuția. Definind o cale q, prin laticea care include cuvântul w, în secvența de cuvinte W, fiind dată secvența de observație X, conduce la expresia probabilității cumulative în latice:

$$p(q, X, W) = p(X \mid q)^{\frac{1}{\gamma}} P(W), \tag{1}$$

unde factorul γ este utilizat pentru a reduce scala modelului acustic de probabilități; valoarea sa este specifică pentru sistem și se încadrează în intervalul 6 - 20. Graficul probabilității a posteriori a unui arc α poate fi estimat prin însumarea tuturor căilor din latice care trec prin acest arc Q_α:

$$p(\alpha \mid X) = \frac{\sum_{q \in Q_\alpha} p(q, A)}{p(X)} \tag{2}$$

Cumularea căilor laticei poate fi calculată eficient folosind algoritmi specifici. Laticea conține arce multiple care au aceeași identitate a cuvintelor, dar corespund unor segmentări temporale ușor diferite și unor contexte n-gram. În cazul estimării a posteriori la nivel de cuvânt, segmentarea și contextul nu sunt oricum de o importanţă semnificativă. Următoarea etapă a tehnicii implică așadar agregarea arcelor corespondente aceluiași cuvânt în enunț. Determinarea arcelor ce trebuie agregate reprezintă problema principală a acestui algoritm precum și sursa principală generatoare de erori. Soluții potențiale de agregare au fost prezentate de Wessel et al. (1998), Wessel et al. (2001), unde o distribuție posteriori de cuvinte este definită într-un interval de timp specific corespunzând aceluiași cuvânt. Valorile probabilităților a posteriori sunt apoi calculate pe baza sumei dintre seturile de arce considerate parte a aceleiași ferestre de timp. Una dintre primele metode propuse a fost C_{sec}, în care se însumează toate arcele cu aceeași identitate de cuvânt care se suprapun cu totul arcului curent. C_{med} restricționează suma acestor arce care au aceeași identitate de cuvânt dar se suprapun cu mediana ferestrei de timp a arcului curent. C_{max} funcționează pe același principiu ca C_{sec}, diferența constând în faptul că în loc să acumuleze probabilitățile posteriori, doar cea mai are valoare posteriori a arcului suprapus este selectată ca și scor. Metoda C_{max} a arătat cele mai bune valori posteriori în contextul utilizării lor ca și scoruri de încredere Wessel et al. (1998), Wessel et al. (2001).

2.2.2 Rețele de Confuzie

Rețelele de confuzie sunt o alternativă, o reprezentare compactă a celor mai posibile ipoteze dintr-un lanț. Toate căile într-o rețea de confuzie sunt constrânse să treacă prin toate nodurile, rezultând într-o simplă reprezentare grafică a acestor ipoteze. Arcele într-o rețea de confuzie corespund cuvintelor (arcele de tip ϵ sunt adăugate pentru cuvinte lipsă din ipoteză). Nodurile impun în esență o segmentare a cuvântului în seturi de confuzie. Figura 1 prezintă un exemplu tipic de rețea de confuzie. Un algoritm de generare a unei rețele de confuzie este descrisă în Mangu et al. (2000) și se bazează pe următorul principiu:

- Probabilitatea a posteriori pentru fiecare arc din lanț este calculată folosind Ecuația (2)
- Probabilitatea a posteriori pentru fiecare cuvânt din lanț, cu un timp de început t_s și un timp de final t_e este calculată folosind:

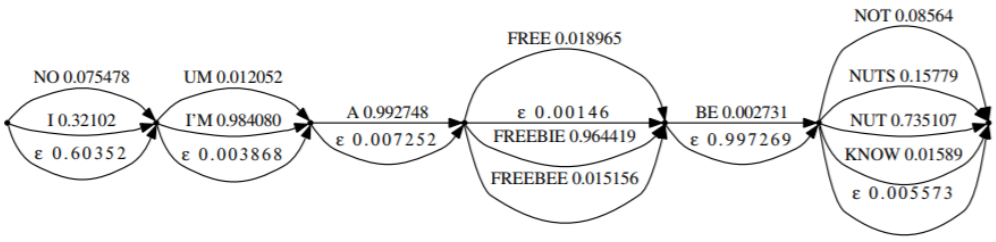

Figura 1: Rețea de confuzie

$$p(W \mid t_s, t_e, X) \tag{3}$$

- Gruparea intra cuvânt. Grupurile ce corespund aceluiași cuvânt și se suprapun în timp sunt fuzionate. Deciziile eronate de grupare în acest punct pot avea un impact negativ semnificativ pentru deciziile de grupare luate ulterior în cadrul procesului. Pentru a evita pe cât posibil acest lucru, posibilele elemente ale unui grup sunt procesate în ordinea valorii a posteriori, calculată pe baza nivelului de suprapunere între ele.

- Grupare între cuvinte. Grupurile ce corespund diferitelor cuvinte care se consideră că aparțin aceluiași set de confuzie sunt fuzionate în acest pas. Cuvintele sunt adăugate în grupuri pe baza timpului de suprapunere și scorul de similaritate fonetică dat de probabilitatea a posteriori.

- În final, arcele ϵ ce corespund cuvintelor nule sunt adăugate în graf. Probabilitatea a posteriori a acestui arc este probabilitatea rămasă astfel încât suma tuturor probabilităților a posteriori trebuie să fie 1.

Scorurile de încredere estimate în această manieră sunt folosite în sistemul RAV pentru a minimiza Word Error Rate (WER) în diferite sarcini recunoaștere vocală. Evermann and Woodland (2000) utilizează acest algoritm pe o bază de date de tip Conversație la telefon (CTS - Conversational Telephone Speech).

2.2.3 Dezavantaje

Aceasta metodă tinde să supraestimeze adevărata distribuție a posteriori, din moment ce probabilitățile sunt calculate pe un subspațiu al ipotezei. Rețelele de confuzie sunt de asemenea susceptibile la ipotezele de independenţă făcute de sistem. Valorile brute ale probabilităților a posteriori trebuie astfel să fie translatate într-un scor de încredere după un anumit algoritm (cum au prezentat și Evermann and Woodland (2000)). Aceasta necesită implementarea unui nou pas după procesare, înainte ca scorurile să fie asignate cuvintelor din transcriere. Mai mult, natura euristică a grupării arcelor și abordările consensuale de grupare bazate pe scoruri de suprapuneri și similaritate nu sunt ideale, și reprezintă o potențială sursă de eroare în procesul de estimare al probabilității posteriori pentru cuvinte.

2.3 Abordarea bazată pe verificarea enunțurilor

Similar cu modul în care este tratată sarcina de verificare a vorbitorului, estimarea încrederii poate fi formulată ca o sarcină de verificare a enunțurilor (VE). Verificarea enunțurilor presupune, similar cu abordarea bazată pe clasificare, atribuirea unui scor de încredere transcrierilor generate de sistemul RAV. Printre primele lucrări publicate folosind această abordare a fost Rose et al. (1995), articol în care VE se aplică în cadrul unui sistem de detecție a cuvintelor cheie. În cele ce urmează vom descrie, pe scurt, această abordare.

Dat fiind un mesaj vorbit X, transcris de un sistem RAV ca un cuvânt W pe baza modelului HMM λ_w, sarcina de a decide dacă rezultatul este corect sau nu poate fi formulată ca o ipoteză într-o problemă de testare. Ipoteza nulă H_0, precum și ipoteza alternativă H_1 sunt definite precum urmează:

- H_0: X a fost corect transcris și este generat de HMM λ_W
- H_1: X a fost incorect transcris și este generat de HMM λ_A,

unde λ_A este un HMM ce corespunde tuturor ipotezelor alternative de cuvinte care sunt, bineînțeles eronate.

Cea mai populară metodă de a respinge o ipoteză în favoarea alteia este testul raportului de maximă plauzibilitate. Modelând H_0 și H_1 cu λ_W și λ_A, fiecare distribuție poate fi evaluată folosind raportul de plauzibilitate:

$$LR(X, \Lambda_W, \Lambda_A) = \frac{P(X \mid \Lambda_W)}{P(X \mid \Lambda_A)} \underset{H_0}{\overset{H_1}{\gtrless}} \tau \tag{4}$$

Decizia de a accepta sau a respinge ipoteza H_0 este luată în funcție de o valoare de prag (τ). Diferența dintre valoarea LR și valoarea pragului poate fi astfel utilizată ca scor de încredere Lleida and Rose (1996). Principala provocare a acestei abordări constă în estimarea modelului alternativ λ_A, numit model de fundal sau anti-model. Studiile anterioare au arătat că aceste modele ar trebui să aibă aceeași structură HMM cu modelul corect λ_W și că ar trebui antrenate discriminativ, folosind criterii precum eroarea de clasificare minimă sau eroare de verificare minimă Rose et al. (1995), Sukkar et al. (1996), Rahim et al. (1997).

Niciuna dintre lucrările menționate mai sus nu a rezolvat complet problema definirii și estimării modelelor alternative, astfel că această abordare de estimare a încrederii a fost abandonată la începutul anilor 2000.

3 Metode de evaluare

Această secțiune discută evaluarea calității metodelor de estimare a scorurilor de încredere (ESI). Deși metodele de ESI sunt utilizate adesea ca o parte a unui sistem mai amplu, a cărui performanță este cea care ne interesează în final, putem totuși evalua separat sarcina de ESI. Intuitiv, spunem că o metodă de ESI funcționează bine dacă scorurile se corelează cu corectitudinea transcrierii: pentru scoruri mari obținem transcrieri corecte, iar pentru scoruri mici, transcrieri incorecte. Pentru evaluare, pentru fiecare cuvânt transcris avem un scor de încredere și o decizie binară referitoare la transcriere – a fost sau nu transcris corect. Notăm scorul de încredere ca un cuvânt să fi fost transcris corect $\hat{p}$ și presupunem că $\hat{p} \in [0, 1]$; căciulița $\hat{\cdot}$ denotă faptul că $\hat{p}$ este un estimat al probabilității p, ca un cuvânt să fie transcris corect.

Pentru acest tip de scenariu, se folosesc metrici întâlnite în comunitatea de *information retrieval*. Pe baza unui prag τ trecem de la un scor de încredere la o valoare binară (predicția ca un cuvânt să fi fost transcris corect sau nu), și definim următoarele metrici: *true positives* (TP), *true negatives* (TN), *false positives* (FP), *false negatives* (FN). Tabelul următor sumarizează calculul fiecăreia dintre metricile enumerate. Spre exemplu, *true positives* este numărul de cuvinte pentru care *(i)* transcrierea este corectă și *(ii)* scorul este mai mare decât pragul impus, τ.

		transcriere corectă	transcriere greșită
scor	$\geq \tau$	TP	FP
	$\leq \tau$	FN	TN

Dacă baleiem pragul $\tau \in [0, 1]$, obținem grafice care exemplifică compromisul între diferitele tipuri de metrici:

- *Receiver operator curve* (ROC) compară rata de valori *true positive* (TPR) cu rata de valori *false positive* (FPR). Mai precis, TPR este raportul dintre TP și numărul de elemente pozitive (numărul de transcrieri corecte), iar FPR este raportul dintre FP și numărul de elemente negative (numărul de transcrieri incorecte). O metodă de ESI funcționează mai bine cu cât TPR-ul este mai mare, iar FPR-ul mai mic. Pragul τ poartă numele de "punct de operare" – de unde și denumirea metodei. Pentru $\tau = 0$, atât TPR, cât și FPR sunt 1, iar pentru $\tau = 1$, ambele metrici ating valoarea de 0.

- *Detection error tradeoff* (DET) compară ratele celor două tipuri de erori: rata de valori *false negative* (FNR) și rata de valori *false positive* (FPR). FNR este raportul dintre FN și numărul de elemente pozitive și poartă numele de "detecții pierdute" (en., *missed detections*); FPR este raportul dintre FP și numărul de elemente negative și poartă numele de "alarme false" (en. *false alarms*). Pentru $\tau = 0$, FNR este 0, iar FPR 1, iar pentru $\tau = 1$, FNR este 1, iar FPR 0. Pentru a raporta un singur număr, de obicei, se folosește pragul τ pentru care FPR este egal cu FNR; această eroare poartă numele de *equal error rate* (EEQ).

Un alt tip de metrică, introdusă de NIST și inspirată din teoria informației, este cross-entropia normalizată (en., *normalized cross-entropy*; NCE). Intuitiv această metrică măsoară diferența între două distribuții de probabilitate: în cazul nostru, distribuția scorurilor de încredere și distribuția corectitudinii cuvintelor. O entropie mică înseamnă că putem spune dacă un cuvânt a fost transcris corect sau nu, o entropie mare se traduce prin incertitudine mare; în consecința dorim un NCE cât mai mic (obținem un NCE nul atunci când cele două distribuții de probabilitate sunt identice).

Concret, formula de calcul a NCE este:

$$\text{NCE} = \frac{H(C) - H(C|X)}{H(C)}, \tag{5}$$

Articol	ER	ROC	DET	NCE
(Wessel et al., 1998)	•	–	–	–
(Wessel et al., 2001)	•	–	•	–
(Hazen et al., 2002)	–	•	–	–
(Fabian, 2007)	•	•	–	–
(Seigel, 2013)	–	–	–	•
(Huang et al., 2013a)	•	–	–	–
(Kalgaonkar et al., 2015)	–	•	–	–
(Cortina, 2016)	•	•	–	•
(Ogawa and Hori, 2017)	•	–	–	•
(Del-Agua et al., 2018)	•	•	–	•

Tabelul 1: Metrici de evaluare utilizate în literatură. Pentru fiecare articol indicăm care metrici sunt folosite pentru evaluare: *error rate* (ER), *receiver operator curve* (ROC), *detection error tradeoff* (DET), *normalized cross-entropy* (NCE). Pe lângă metricile enumerate în table, Huang et al. (2013a) folosesc divergența Kulback-Leibler, Cortina (2016) folosește h-measure, iar Ogawa and Hori (2017) folosesc F-measure.

unde $H(-)$ indică entropia unei variabile aleatoare. Dacă considerăm n ca numărul de cuvinte, iar m numărul de cuvinte ce au fost transcrise corect, atunci cantitățile din ecuația 5 sunt definite astfel:

$$H(C) = -m \log \frac{m}{n} - (n-m) \log \frac{n-m}{n} \tag{6}$$

$$H(C|X) = \sum_{c_i=1} -\log \hat{p} + \sum_{c_i=0} -\log(1-\hat{p}). \tag{7}$$

Deși comună în evaluarea metodelor de ESI (vezi tabelul 1), metrica NCE a fost criticată de-a lungul timpului. Wessel et al. (1998) menționează că nu este definită atunci când avem cuvinte pentru care probabilitatea estimată este 0 sau 1. De fapt, NIST menționează că uneltele lor evită astfel de cazuri înlocuind scorurile de 0 și 1 cu 10^{-7}, respectiv $1 - 10^{-7}$. Siu and Gish (1999) observă că NCE are dezavantajul de a fi sensibilă la schimbări în performanța sistemului de transcriere. Ei oferă ca alternativă metrica *equal error rate* (EEQ): este un singur număr, intuitivă, ușor de calculat și robustă la schimbări în probabilitățile de transcriere corectă. Tabelul 1 sumarizează metricile de evaluare utilizate în literatură.

4 Auto-antrenarea modelelor acustice pentru RAV utilizând scoruri de încredere

Una dintre cele mai importante aplicații ale estimării încrederii pentru transcrierile rezultate în urma procesului de recunoaștere automată a vorbirii (RAV) este auto-antrenarea (en., self-training) modelelor acustice pentru RAV. Scopul acestui proces este bineînțeles creșterea performanțelor acestor sisteme. Dat fiind faptul că, indiferent de limbă, mai multe date de antrenare constituie de regulă premisele obținerii unui model acustic de RAV mai performant, astfel de aplicații sunt intens utilizate atât la nivel științific, dar mai ales comercial. Procesul de auto-antrenare are, în general, următoarele etape:

1. Antrenarea unui model acustic "sămânță" (en. seed model) folosind un corpus de vorbire adnotat manual.
2. Crearea unui sistem de RAV folosind modelul acustic sămânță.
3. Transcrierea unui corpus de vorbire neadnotată folosind sistemul RAV de la punctul anterior.
4. Selecția unei porțiuni din corpusul transcris automat pentru care încrederea privind corectitudinea transcrierii este mare.
5. Reantrenarea modelului acustic sămânță cu vorbirea adnotată manual și noul corpus rezultat la punctul anterior.

În acest context, estimarea încrederii este necesară la punctul 4. De obicei se utilizează scoruri de încredere la nivel de cadru, cuvânt sau propoziție.

Veselỳ et al. (2013) aplică această metodologie pentru crearea unui sistem de RAV pentru limba vietnameză. Autorii pornesc de la un model acustic de tip rețea neurală profundă (en. deep neural network - DNN) antrenat pe corpusul de antrenare din sarcina "RAV pentru limbă surpriză" din programul IARPA Babel. Antrenarea modelului acustic sămânță (pasul 1) s-a realizat pe corpusul de antrenare din condițiile LLP (en. limited language pack), adică 10.8 ore de vorbire, iar evaluarea metodologiei (3, 4, 5) s-a realizat utilizând corpusul din scenariul FLP (en. full language pack): 74 ore de vorbire. Din punctul de vedere al scorurilor de încredere, Veselỳ et al. (2013) utilizează (i) scorul de încredere la nivel de propoziție, calculat ca media a scorurilor de încredere la nivel de cuvânt și (ii) scorul de încredere la nivel de cadru.

Concluziile autorilor sunt că (i) scorurile de încredere la nivel de cuvânt sau propoziție nu sunt utile (filtrarea transcrierilor obținute la pasul 3 în funcție de aceste scoruri nu ajută la nimic), (ii) se obține o scădere absolută a WER de 1.4% (de la 63.1% la 61.7%) reantrenând modelul acustic (pasul 5) cu toate transcrierile obținute automat și (iii) o scădere suplimentară de 0.8% a WER reantrenând modelul acustic cu transcrieri filtrate pe baza scorului de încredere la nivel de cadru.

Walker et al. (2017) utilizează metodologia standard prezentată mai sus și descriu în detaliu procesul prin care au antrenat un sistem de RAV pentru limba engleză. Autorii pornesc de la un model acustic de tip DNN-HMM antrenat cu aproximativ 2k ore de vorbire conversațională (folosind Kaldi nnet2) și un model de limbă antrenat cu corpusul Fisher English (folosind SRI-LM). Corpusul de vorbire neadnotată transcris automat (pasul 3) cuprinde aproximativ 25k ore de vorbire telefonică. Autorii folosesc ideea introdusă în (Kapralova et al., 2014) de a utiliza la acest pas un sistem RAV mai performant, dar mai lent (cu model de limbă amplu și beam width mare) pentru a genera o latice de căutare mai amplă în vederea generării unor scoruri de încredere mai bune. Autorii propun selecția porțiunilor transcrise automat pe baza unui scor de încredere la nivel de propoziție, numit Minimum Bayes Risk (MBR) (Xu et al., 2011), și perplexității la nivel de propoziție date de un model de limbă mare (5-gram, 125k cuvinte, 5M n-grame). Lucrarea arată că scorul de încredere MBR (generat cu Kaldi lattice-mbr-decode) se corelează foarte bine cu WER-ul propoziției în cauză, spre diferență de scorul de încredere la nivel de propoziție obținut prin diferența dintre costul total al celei mai bune căi și "a doua cea mai bună cale" din laticea rezultat (generat cu Kaldi lattice-confidence).

În urma aplicării metodologiei, autorii reușesc să selecteze 5k ore de vorbire transcrisă din cele 25k ore inițiale cu care reantrenează un model acustic mai complex (6 straturi ascunse dens conectate, față de 4 în modelul sămânță). De asemenea, autorii propun reantrenarea modelului de limbă folosind transcrierile selectate automat. Lucrarea raportează că prin aplicarea acestei metodologii s-a obținut o scădere relativă a WER de 35% (de la 22.1 la 14.27) pentru vorbirea operatorului, respectiv de 19% (de la 21.6 la 17.5) pentru vorbirea provenind de la client.

Kapralova et al. (2014) aplică aceeași metodologie de mai sus pentru a genera sisteme de RAV mai performante pentru sarcina de căutare vocală a Google (în mai multe limbi: rusă, franceză și portugheză). Autorii folosesc pentru filtrarea transcrierilor generate automat un scor de încredere la nivel

de propoziție derivat din probabilitățile a posteriori ale cuvintelor existente în laticea rezultat. Ei arată că acest scor de încredere se corelează bine cu WER-ul propoziției și că în general primele 10% din propozițiile sortate în funcție de scorul de încredere au un WER sub 10%. Un alt lucru specific în acest studiu este limitarea numărului de transcrieri generate automat (cu scor de încredere mare) la o valoare maximă N (20 în cazul acesta). Altfel spus, autorii utilizează pentru reantrenarea modelului acustic maxim N transcrieri identice (dintre cele generate automat). Motivația este de a obține date care să acopere mai bine repertoriul fonetic și un vocabular mai mare pentru transcrierile rezultate automat.

Huang et al. (2013b) utilizează de asemenea scoruri de încredere la nivel de cuvânt pentru realizarea selecției datelor ce urmează a fi folosite pentru reantrenare, însă aplică o procedură de selecție mai complexă. Mai concret, autorii folosesc trei sisteme de RAV distincte (cu modele acustice și lingvistice distincte) pentru a transcrie vorbirea neadnotată (pasul 3) și apoi o procedură standard de combinare a rezultatelor (laticelor de transcrieri) numită ROVER (Fiscus, 1997) obținând astfel o latice mai bună. În continuare autorii propun utilizarea unui sistem de votare de tip comitet pentru recalibrarea scorurilor de încredere din laticea generată cu sistemul ROVER și folosesc această latice pentru a selecta propoziții cu scor de încredere mare (pasul 4). Pentru experimentele prezentate în lucrare se selectează pentru reantrenare primele 60% din totalul propozițiilor transcrise automat, în ordinea scorurilor de încredere, dar fără a se pune vreun prag pe valoarea absolută a scorului de încredere.

Utilizând această metodologie, autorii raportează o scădere relativă a WER de 7.2% pentru un sistem de RAV de tip HMM-GMM antrenat discriminativ (după adăugarea a 2.1k ore de vorbire ne-transcrisă), de 28.2% pentru un sistem de RAV de tip HMM-GMM neadaptat la domeniu antrenat generativ (după adăugarea a 10k ore de vorbire ne-transcrisă), respectiv de 15.0% pentru un sistem de RAV de tip DNN-HMM (după adăugarea a 1k ore de vorbire ne-transcrisă). Toate experimentele au fost realizate pe limba engleză pe o sarcină de transcriere de mesaje dictate scurte.

Koctúr et al. (2016) aplică de asemenea o metodologie hibridă de obținere automată a transcrierilor necesare reantrenării sistemelor de RAV. Autorii folosesc două sisteme complementare de RAV pentru a obține două transcrieri ale datelor de intrare, selectează porțiunile transcrise identic și aplică pe acestea o filtrare (cu prag variabil) bazată pe scoruri de încredere la nivel de cuvânt. Inovația constă în faptul că pragul aplicat pe scorul de încredere la

nivel de cuvânt descrește cu fiecare cuvânt consecutiv dintr-o serie mai lungă de cuvinte aliniate.

Jouvet and Fohr (2014) utilizează și ei o metodologie similară cu cea descrisă în (Koctúr et al., 2016), însă ajung la concluzia că aplicarea unui scor de încredere nu ajută atât de mult. Cele mai bune rezultate raportate în această lucrare sunt ale un sistem ce folosește strict transcrierile aliniate obținute în urma decodării cu două sisteme RAV complementare.

Metoda propusă de Su and Xu (2015) se bazează pe ipoteza că alinierea datelor ne-transcrise nu este precisă și ea este cea care deteriorează performanța modelelor auto-antrenate. Astfel ideea lor este ca auto-antrenarea să nu altereze ultimul strat al DNN-ului, despre care autorii spun că este mai sensibil la alinieri. Arhitectura propusă folosește două capete: *(i)* un strat pentru predicția senonelor din baza de date sămânță; *(ii)* un strat pentru predicția senonelor din baza de date neadnotată, care sunt auto-transcrise. Antrenarea se face simultan pentru ambele tipuri de date, iar la final se încheie cu un pas de reantrenare al primului capăt, care va fi utilizat pentru evaluare. Su and Xu (2015) observă că dacă ar fi folosească metoda standard de auto-antrenare pe cele mai confidente 20% din date, performanța nu se îmbunătățește; în cazul metodei lor obțin o performanță ceva mai bună, care surpinzător este atinsă și dacă folosesc cele mai puțin confidente 20% din date. Acest rezultat dovedește că scorurile de încredere, deși sunt corelate cu WER-ul (după cum este prezentat în articol), nu aduc neapărat o performanță mai bună a metodelor de auto-antrenare. Autorii folosesc o metodă de estimare a scorurilor de încredere bazată pe probabilitatea a posteriori (din SCLite).

5 Concluzii

Această lucrare conturează starea artei în domeniul estimării încrederii sistemelor de recunoaștere automată a vorbirii folosind scoruri de încredere. EI prezintă o importanță majoră în îmbunătățirea sistemelor de RAV, contribuind la procesul de detecție și corectare a erorilor. Metodele de generare a scorurilor de încredere se încadrează în trei mari categorii: abordări bazate pe clasificare, pe probabilități posteriori și pe verificarea enunțurilor, ce au fost detaliate în Secțiunea 2. Metodele principale de evaluare a scorurilor de încredere: curbele ROC și DET precum și NCE au fost prezentate în Secțiunea 3. Ulterior, Secțiunea 4 descrie una dintre cele mai importante aplicații

ale EI și anume auto-antrenarea modelelor acustice pentru RAV.

Dintre toate metodele de EI discutate în această lucrare, cele mai slabe rezultate au fost înregistrate în abordările bazate pe probabilitățile aposteriori, întrucât metodele actuale nu reușesc să estimeze corect distribuția aposteriori. Cele mai bune performanțe au fost obținute în contextul utilizării unor trăsături predictive, ce pot fi extrase din laticea de recunoaștere, din modelul de limbă sau din informația acustică propriu-zisă. Aceste trăsături sunt ulterior introduse la intrarea a unui clasificator. În literatura de specialitate au fost propuși o serie de clasificatori, dintre care s-au remarcat sistemele bazate pe CRF și cele bazate pe RNN-uri. Cea mai promițătoare direcție de dezvoltare în prezent o reprezintă utilizarea BiDRNN-urilor, ce folosesc în procesul de decizie un context bidirecțional. Sistemele BiDRNN au depășit performanțele celor bazate pe CRF, demonstrând o putere mai mare de generalizare.

Bibliografie

Chase, L. L. (1997). *Error-responsive feedback mechanisms for speech recognizers*.

Cortina, I. S. (2016). *Confidence Measures for Automatic and Interactive Speech Recognition*. PhD thesis, Universitat Politècnica de València.

Del-Agua, M. A., Gimenez, A., Sanchis, A., Civera, J., and Juan, A. (2018). Speaker-adapted confidence measures for ASR using deep bidirectional recurrent neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, PP(99):1–1.

Evermann, G. and Woodland, P. C. (2000). Large vocabulary decoding and confidence estimation using word posterior probabilities. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 3, pages 1655–1658. IEEE.

Fabian, T. (2007). *Confidence measurement techniques in automatic speech recognition and dialog management*. PhD thesis, Technical University München.

Fayolle, J., Moreau, F., Raymond, C., Gravier, G., and Gros, P. (2010). Crf-based combination of contextual features to improve a posteriori word-level confidence measures. In *Eleventh Annual Conference of the International Speech Communication Association*.

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.

Goel, V. and Byrne, W. J. (2000). Minimum bayes-risk automatic speech recognition. *Computer Speech & Language*, 14(2):115–135.

Hazen, T. J., Seneff, S., and Polifroni, J. (2002). Recognition confidence scoring and its use in speech understanding systems. *Computer Speech & Language*, 16(1):49–67.

Hori, T., Kubo, Y., and Nakamura, A. (2014). Real-time one-pass decoding with recurrent neural network language model for speech recognition. In

Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on, pages 6364–6368. IEEE.

Huang, P.-S., Kumar, K., Liu, C., Gong, Y., and Deng, L. (2013a). Predicting speech recognition confidence using deep learning with word identity and score features. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7413–7417. IEEE.

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

Jiang, H. (2005). Confidence measures for speech recognition: A survey. *Speech communication*, 45(4):455–470.

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.

Kalgaonkar, K., Liu, C., Gong, Y., and Yao, K. (2015). Estimating confidence scores on ASR results using recurrent neural networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4999–5003. IEEE.

Kamppari, S. O. and Hazen, T. J. (2000). Word and phone level acoustic confidence scoring. In *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, volume 3, pages 1799–1802. IEEE.

Kapralova, O., Alex, J., Weinstein, E., Moreno, P. J., and Siohan, O. (2014). A big data approach to acoustic model training corpus selection. In *INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association, Singapore, September 14-18, 2014*, pages 2083–2087.

Kemp, T. and Schaaf, T. (1997). Estimating confidence using word lattices. In *Eurospeech*.

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.

Lleida, E. and Rose, R. C. (1996). Likelihood ratio decoding and confidence measures for continuous speech recognition. In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, volume 1, pages 478–481. IEEE.

Mangu, L., Brill, E., and Stolcke, A. (2000). Finding consensus in speech recognition: word error minimization and other applications of confusion networks. *Computer Speech & Language*, 14(4):373–400.

Marin, A., Kwiatkowski, T., Ostendorf, M., and Zettlemoyer, L. (2012). Using syntactic and confusion network structure for out-of-vocabulary word detection. In *Spoken Language Technology Workshop (SLT), 2012 IEEE*, pages 159–164. IEEE.

Mikolov, T., Deoras, A., Povey, D., Burget, L., and Černockỳ, J. (2011). Strategies for training large scale neural network language models. In *IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 196–201. IEEE.

Mikolov, T., Karafiát, M., Burget, L., Černockỳ, J., and Khudanpur, S. (2010). Recurrent neural network based language model. In *Interspeech*, volume 2, page 3.

Neti, C. V., Roukos, S., and Eide, E. (1997). Word-based confidence measures as a guide for stack search in speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 2, pages 883–886. IEEE.

Ogawa, A. and Hori, T. (2017). Error detection and accuracy estimation in automatic speech recognition using deep bidirectional recurrent neural networks. *Speech Communication*, 89:70–83.

Pinto, J. and Sitaram, R. (2005). Confidence measures in speech recognition based on probability distribution of likelihoods. In *Ninth European Conference on Speech Communication and Technology*.

Rahim, M. G., Lee, C.-H., and Juang, B.-H. (1997). Discriminative utterance verification for connected digits recognition. *IEEE transactions on speech and audio processing*, 5(3):266–277.

Rose, R. C., Juang, B.-H., and Lee, C.-H. (1995). A training procedure for verifying string hypotheses in continuous speech recognition. In *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, volume 1, pages 281–284. IEEE.

Rueber, B. (1997). Obtaining confidence measures from sentence probabilities. In *Eurospeech*.

Seigel, M. S. (2013). *Confidence estimation for automatic speech recognition hypotheses*. PhD thesis, University of Cambridge.

Siu, M. and Gish, H. (1999). Evaluation of word confidence for speech recognition systems. *Computer Speech & Language*, 13(4):299–319.

Su, H. and Xu, H. (2015). Multi-softmax deep neural network for semi-supervised training. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Sukkar, R. A., Setlur, A. R., Rahim, M. G., and Lee, C.-H. (1996). Utterance verification of keyword strings using word-based minimum verification error (wb-mve) training. In *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, volume 1, pages 518–521. IEEE.

Veselỳ, K., Hannemann, M., and Burget, L. (2013). Semi-supervised training of deep neural networks. In *Workshop on Automatic Speech Recognition and Understanding*, pages 267–272. IEEE.

Walker, S., Pedersen, M., Orife, I., and Flaks, J. (2017). Semi-supervised model training for unbounded conversational speech recognition. *CoRR*, abs/1705.09724.

Weintraub, M., Beaufays, F., Rivlin, Z., Konig, Y., and Stolcke, A. (1997). Neural-network based measures of confidence for word recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 2, pages 887–890. IEEE.

Wessel, F., Macherey, K., and Schluter, R. (1998). Using word probabilities as confidence measures. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages 225–228. IEEE.

Wessel, F., Schluter, R., Macherey, K., and Ney, H. (2001). Confidence measures for large vocabulary continuous speech recognition. *IEEE Transactions on Speech and Audio Processing*, 9(3):288–298.

Willett, D., Worm, A., Neukirchen, C., and Rigoll, G. (1998). Confidence measures for hmm-based speech recognition. In *Fifth International Conference on Spoken Language Processing*.

Xu, H., Povey, D., Mangu, L., and Zhu, J. (2010). An improved consensus-like method for minimum bayes risk decoding and lattice combination. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 4938–4941. IEEE.

Xu, H., Povey, D., Mangu, L., and Zhu, J. (2011). Minimum bayes risk decoding and system combination based on a recursion for edit distance. *Computer Speech and Language*, 25(4):802 – 828.

Young, S. J. and Woodland, P. C. (1994). State clustering in hidden markov model-based continuous speech recognition. *Computer Speech & Language*, 8(4):369–383.

Yu, D., Wang, S., Li, J., and Deng, L. (2010). Word confidence calibration using a maximum entropy model with constraints on confidence and word distributions. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 4446–4449. IEEE.

Zhang, R. and Rudnicky, A. I. (2001). Word level confidence annotation using combinations of features.