Vorbis: Recunoașterea automată a vorbirii din semnale multimodale

Raport tehnico-științific final

Dan Oneață
Universitatea POLITEHNICA din București
dan.oneata@speed.pub.ro

Cuprins

1 Obiectivele proiectului

Scopul proiectului Vorbis l-a reprezentat sarcina recunoașterii automate a vorbirii pe baza semnalelor multimodale, și anume semnale audio și semnale vizuale. Pentru a atinge acest obiectiv am tratat adițional două obiective intermediare legate de fiecare din cele două modalități (audio și vizuală). Aceste trei obiective țin de domeniul învățării automate și inteligenței artificiale și, concret, sunt următoarele:

O1 Recunoașterea automată a vorbirii folosind învățarea de tip end-to-end;

O2 Înțelegerea informației vizuale din imagini folosind învățarea de tip end-to-end;

O3 Recunoașterea automată a vorbirii din semnale multimodale prin fuzionarea celor două tipuri arhitecturi de tip end-to-end (O1 și O2).

Primele două obiective, O1 și O2, au făcut subiectul etapei 2, iar obiectivul O3 a fost realizat în etapa 3. Reamintim că rapoartele precedente și informații adiționale privind proiectul sunt disponibile pe pagina proiectului: http://vorbis.speed.pub.ro/

2 Rezultate și activități

Conform planului de realizare am avut de livrat următoarele rezultate în cadrul proiectului:

R1.1 Studiu asupra stării artei.

R2.1 Sistem de recunoaștere automată a vorbirii de tip end-to-end.

R2.2 Sistem pentru înțelegerea informației vizuale de tip end-to-end.

R2.3 Un articol de conferință.

R3.1 Sistem multimodal de tip end-to-end pentru recunoașterea automată a vorbirii.

R3.2 Un articol de conferință și un articol de jurnal.

Aceste rezultate au fost atinse în proporție de 100% și au fost realizate prin activitățile definite în planul de realizare:

A1.1 *Studiu asupra stării artei.* În această activitate am conspectat și sumarizat cele mai relevante articole pentru sarcina de recunoaștere automată a vorbirii din semnale multimodale. Literatura de specialitate a indicat faptul că aceasta este o direcție în continuă explorare, numai în ultimii doi ani (de la momentul scrierii raportului 1) fiind publicate șapte articole foarte relevante pe această temă. Am observat că tendința este de a utiliza rețele de tip end-to-end care integrează ambele semnale (audio și imagini) la mai multe nivele și încercă obținerea de adnotări cât mai structurate între text și intrarea vizuală. În final, am plasat sarcina de interes în contextul altor sarcini multi-modale și am identificat multiple direcții de cercetare în contextul lucrărilor recent publicate. Acest studiu este disponibil pe pagina proiectului.

A2.1 *Proiectarea și implementarea unui sistem de recunoaștere automată a vorbirii de tip end-to-end.* În cadrul acestei activități am dezvoltat un sistem de recunoaștere automată a vorbirii (RAV) la nivelul stării artei bazat pe arhitectura de tip Transformer [11, 39, 43] și implementat cu ajutorul utilitarului ESPnet [40]. Am adus îmbunătățiri sistemului de bază utilizând tehnici precum învățarea prin transfer (pornind de la baza de date LibriSpeech [26]) și augmentarea de date [28]. Toate aceste aspecte au fost validate empiric pe partea audio a trei baze de date populare de tip multi-modal (care conțin triplete de tip audio, imagini, transcrieri). Comparația cu metodele de RAV dezvoltate anterior în cadrul abordărilor multi-modale arată că obținem performanță mult superioară (spre exemplu, pe baza de date How2 [34] atingem o eroare la nivel de cuvânt de 11.2%, față de 17.7% cât a fost raportat anterior de Ghorbani et al. [4]).

A2.2 *Proiectarea și implementarea unui sistem vizual de tip end-to-end.* Pentru această activitate am pornit de la două categorii de modele vizuale: *(i)* arhitectura ResNet [7] pre-antrenată pentru clasificare pe baza de date ImageNet [32] și *(ii)* arhitecturile ResNet și Visual Transformer [3] pre-antrenate în mod auto-supervizat pe o bază de date de 400 de milioane de perechi de imagini și text culese de pe internet [30]. Am extins și evaluat aceste două tipuri de modele pentru două sarcini distincte: înțelegerea semantică a imaginilor și traducerea de la video la audio. Pentru prima sarcină, am arătat că arhitectura de tip Visual Transformer (ViT) oferă capabilități de generalizare remarcabile: chiar dacă nu a fost antrenată explicit pentru această sarcină și nici pe baza de date de interes (Flickr8K [8]), ViT obține rezultate comparabile cu un model specializat. Pentru cea de-a două sarcină, am evaluat arhitectura ResNet într-un scenariu multi-modal, de traducere a semnalului video (mișcarea buzelor) în semnal audio (uteranța rostită). Rezultatele empirice au arătat că ResNet este capabil să capteze informații vizuale fine necesare sintezei audio și chiar și identitatea vorbitorului.

A2.3 *Diseminarea rezultatelor de cercetare prin publicarea unui articol de conferință.* În urma activității de cercetare am publicat următoarele trei articole, dintre care două de revistă (primul și al treilea) și unul de conferință (al doilea); articolele de revistă sunt deja indexate în Web of Science (Thompson Reuters; ISI), iar cel de conferință urmează a fi indexat IEEE Xplore și Web of Science (Thompson Reuters; ISI):

1. D. Oneață and H. Cucu. Multimodal speech recognition for unmanned aerial vehicles. *Computers & Electrical Engineering*, 90:106943, 2021
2. D. Oneață, A. Stan, and H. Cucu. Speaker disentanglement in video-to-speech conversion. In *29th European Signal Processing Conference*, 2021
3. A. Caranica, D. Oneață, H. Cucu, and C. Burileanu. Confidence estimation for lattice-based and lattice-free automatic speech recognition. *UPB Scientific Bulletin, Series C: Electrical Engineering and Computer Science*, 83(3):155–170, 2021

A3.1 *Proiectarea și implementarea unui sistem multimodal de tip end-to-end pentru recunoașterea automată a vorbirii.* În cadrul acestei activități am dezvoltat un sistem de recunoaștere automată a vorbirii (RAV) care utilizează fluxuri de date multimodale (audio și imagini). Sistemul proiectat combină modulele dezvoltate anterior în cadrul proiectului: modulul pentru RAV unimodal, bazat doar pe audio (acesta a fost dezvoltat în cardul activității A2.1), și modulul pentru înțelegerea a informației vizuale (acesta a fost dezvoltat în cadrul activității A2.2). Reprezentările învățate peste cele două fluxuri de date de către cele două module sunt fuzionate prin concatenare și transmise apoi decodorului din sistemul de RAV care produce transcrierea.

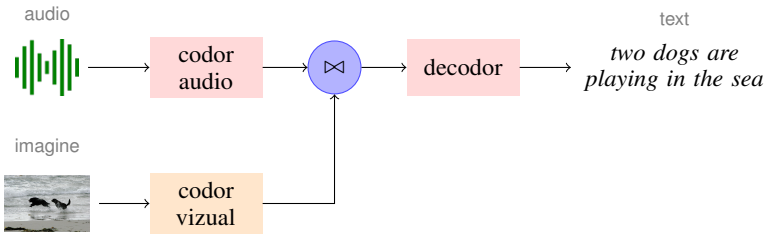


Figura 1: Prezentare generală a sistemului dezvoltat pentru sarcina de recunoaștere automată a vorbirii din semnale multimodale. În comparație cu recunoașterea tradițională a vorbirii, care generează transcripții bazate exclusiv pe intrarea audio, metodele de tip multimodal implică utilizarea unei intrări suplimentare corespunzătoare fluxului vizual (de exemplu, imagine sau video). Motivația principală a acestei configurații este că informațiile vizuale sunt adesea asociate cu audio-ul (așa cum se întâmplă în videoclipuri cu instrucțiuni, documentare, filme) și acestea pot ajuta la dezambiguizarea înregistrării audio, în consecință, producând transcripții mai precise. În acest raport, investigăm modalități de construire a unui sistemul multimodal, concentrându-ne pe cele două codoare (audio și vizual) și fuziunea acestora.

Întreg sistemul este antrenat end-to-end (de la un capăt la altul), rezultatele experimentale confirmând că acest lucru este esențial pentru a obține o performanță optimă (a se vedea Tabelul 3). Implementarea sistemului s-a realizat în limbajul de programare Python și se bazează pe utilitarul ESPnet [40]; codul nostru este disponibil online la adresa: `https://github.com/danoneata/espnet/tree/multimodal-asr/egs2/vorbis`. Comparativ cu starea artei obținem rezultate net superioare: de exemplu, pe baza de date How2 [34] reducem eroarea la nivel de cuvânt la 10.8% față de 17.2%, cât fusese raportat anterior de către Ghorbani *et al.* [4] (a se vedea Secțiunea 3.3.1). De asemenea, din punct de vedere științific am arătat empiric că tehnicile precum învățarea prin transfer (pornind de la baza de date LibriSpeech [26]) și augmentarea de date [28] sunt cruciale mai ales în contextul învățării multimodale (a se vedea Secțiunea 3.3.2).

A3.2 *Diseminarea rezultatelor de cercetare prin publicarea unui articol de conferință și a unui articol de jurnal.* În urma activității de cercetare am publicat următoarele trei articole, dintre care unul de conferință (primul, conferință de rang A) și două de revistă (al doilea și al treilea, ambele Q1). Cel de-al treilea articol de revistă este deja indexat în Web of Science (Thompson Reuters; ISI), iar celelalte conferință urmează a fi indexate de către IEEE Xplore și Web of Science:

1. D. Oneață and H. Cucu. Improving multimodal speech recognition by data augmentation and speech representations. In *Computer Vision and Pattern Recognition Workshops*. IEEE, 2022
2. K. Olaleye, D. Oneață, and H. Kamper. Keyword localisation in untranscribed speech using visually grounded speech models. *IEEE Journal of Selected Topics in Signal Processing*, 2022
3. D. Oneață, B. Lőrincz, A. Stan, and H. Cucu. FlexLip: A controllable text-to-lip system. *Sensors*, 22(11), Jun 2022

3 Descrierea științifică și tehnică

În continuare prezentăm în detaliu aspectele științifice și tehnice ce corespund proiectului. Reamintim că scopul final al proiectului l-a reprezentat sarcina de recunoaștere automată a vorbirii din semnale multimodale. Concret, presupunem că avem două intrări (semnalul acustic și o modalitate vizuală aferentă, cum ar fi un videoclip sau o imagine) și dorim să obținem transcrierea enunțului de intrare. Această configurație este ilustrată vizual în Figura 1.

3.1 Metodologie

În sarcina de recunoaștere automate a vorbirii, intrarea audio $\mathbf{a}$ este mapată la o transcriere $\mathbf{t}$, de obicei reprezentată ca o secvență de jetoane. În abordarea uzuală (de tip codor-decodor) rezultatul este obținut prin compunerea celor două componente: $\mathbf{t} = \mathrm{Dec}(\mathrm{Enc}(\mathbf{a}))$. În cazul recunoașterii multimodale a vorbirii, presupunem că avem acces la o intrare suplimentară—canalul vizual $\mathbf{v}$. Informația vizuală este procesată de un codor separat, $\mathrm{Enc_v}$, și integrată în rețea printr-o funcție de fuziune, pe care o notăm cu "$\bowtie$":

$$\mathbf{t} = \mathrm{Dec}(\mathrm{Enc}(\mathbf{a}) \bowtie \mathrm{Enc_v}(\mathbf{v})).$$

În continuare, discutăm fiecare dintre componente: codorul și decodorul (corespunzătoare sistemului de recunoaștere a vorbirii) în secțiunea 3.1.1 (această secțiune corespunde activității A2.1), codorul vizual în secțiunea 3.1.2 (această

secțiune corespunde activității A2.2), funcția de fuziune în secțiunea 3.1.3 (această secțiune corespunde activității A3.1).

3.1.1 Sistem de recunoaștere automată a vorbirii de tip end-to-end

Baza sistemului multimodal o reprezintă un sistem de recunoaștere automată a vorbirii end-to-end. Folosim o rețea de tip Transformer [39], care se bazează pe module de auto-atenție pentru codor și module de atenție încrucișată în decodor, care agregă informațiile din fluxul audio. Rețeaua prezice jetoane într-o manieră autoregresivă, prin modelarea probabilității următorului jeton dat audio-ul de intrare și jetoanele prezise anterior, $p(t_k|\mathbf{t}_{<k}, \mathbf{a})$. În raportul precedent am explorat două idei de îmbunătățire ale sistemului de recunoaștere a vorbirii, în continuare discutăm cum le extindem în contextul învățării multimodale.

Învățarea prin transfer. Lucrările recente despre recunoașterea multimodală a vorbirii transferă reprezentări vizuale, obținute ca activări sau predicții softmax ale unei rețele de clasificare vizuală preantrenată. În funcție de datele de pregătire (obiecte, ca în [18, 25, 37, 38]; scene, ca în [5, 17, 36]; acțiuni, ca în [1, 17, 27]; fețe, ca în [17, 19]), codorul vizual este antrenat pentru a recunoaște anumite categorii de informații vizuale. Cu toate acestea, niciuna dintre aceste lucrări anterioare nu utilizează reprezentări de vorbire preantrenate. În acest raport nu doar arătăm importanța de a pleca de la o bună reprezentare atât pentru canalele audio cât și pentru cele vizuale, dar, în mod esențial, oferim un răspuns la întrebarea dacă informațiile vizuale sunt utile în cazul unui sistem de bază mai puternic.

Augmentarea datelor. Mărirea setului de antrenare cu date perturbate este o tehnică comună pentru a facilita învățarea invarianților de către modelele neurale adânci, de capacitate mare. Pentru clasificarea imaginilor, imaginile sunt adesea modificate prin răsturnări orizontale și mici transformări afine, în timp ce pentru recunoașterea vorbirii viteza unui enunț este modificată prin deformarea în timp. Folosim aceste idei, în special cele legate de augmentarea datelor de vorbire, pentru a îmbunătăți modelele multimodale. Intuiția noastră este că prin perturbarea semnalului audio modelul va fi încurajat să se bazeze mai mult pe canalul vizual. Inspirația provine din lucrările lui Srinivasan *et al.* [35, 37] care au arătat că modelele multimodale obțin rezultate mai bune comparativ cu sistemul de RAV de bază chiar și atunci când perechile audio-imagine sunt nepotrivite (incongruente) [35]; în schimb, dacă modelele multimodale au fost antrenate pe semnale audio mascate, acest comportament este atenuat [36]. În comparație cu abordările anterioare [35–37], nu ne limităm la mascarea temporală a cuvintelor, ci mascăm aleatoriu segmentele temporale și de frecvență, ca în [28]. În consecință, abordarea noastră este mai generală și mai ușor de utilizat.

3.1.2 Sistem vizual de tip end-to-end

Codorul vizual rezumă informațiile prezente la intrarea canalului vizual—presupunem că avem dată o imagine la intrare. Sistemul nostru se bazează pe arhitectura populară ResNet [7], care a fost folosită și în lucrările anterioare despre recunoașterea multimodală a vorbirii, spre exemplu, [1, 35, 36]. Această arhitectură o inițializăm cu ponderile unui model preantrenat pe setul de date ImageNet [32] și folosim activările intermediare ale rețelei ca reprezentări vizuale în sistemul multimodal. În funcție de stratul din care utilizăm activările, obținem fie *(i)* un singur vector de reprezentare, fie *(ii)* o secvență de vectori de reprezentări. Concret, activările înainte de stratul softmax (și după stratul de agregare de tip medie globală) produc un singur vector de dimensiune fixă, care codifică informații globale din întreaga imagine. În schimb, dacă luăm activările cu un strat înainte (adică înainte de stratul de agregare de tip medie globală), obținem o grilă de dimensiune 7×7, pe care o serializăm într-o listă de $K = 49$ de reprezentări. Această a doua abordare codifică mai multe informații locale, ceea ce sperăm că permite modelului să folosească caracteristici mai fine ale imaginii. Peste secvența de reprezentări, învățăm opțional o rețea perceptron multistrat modulată (eng., *gated multi-layer perceptron*; gMLP) [16]; acest tip de rețea este o arhitectură introdusă recent ca substitut pentru straturile de auto-atenție și care are avantajul că necesită mai puține calcule și memorie, menținând în același timp performanța.

3.1.3 Sistem multimodal pentru recunoașterea automată a vorbirii: Tehnici de fuziune

Tehnicile de fuziune propuse combină reprezentările audio și vizuale (produse de fiecare dintre cele două codoare). Odată fuzionate într-o reprezentare comună, aceasta este apoi trimisă către decodor care produce transcripția corespunzătoare. În continuare presupunem că reprezentările audio au dimensiunea $D_{\mathbf{a}} \times T$, în timp ce cele vizuale au dimensiunea $D_{\mathbf{v}} \times K$ (a doua axă, de lungime K, poate corespunde unei liste de detecții dintr-o imagine sau unei liste de cadre dintr-un videoclip). Experimentăm două abordări de fuziune, fie de-a lungul dimensiunii reprezentării (`emb`, de la *embedding*), fie de-a lungul dimensiunilor secvenței (`seq`, de la *sequence*); aceste două variante sunt ilustrate în Figura 2. Alegerea tehnicii de fuziune depinde și de codorul vizual: dacă reprezentăm intrarea vizuală cu un singur vector caracteristic ($K = 1$) putem doar concatena de-a lungul dimensiunii reprezentării (cea de dimensiune D), în timp ce

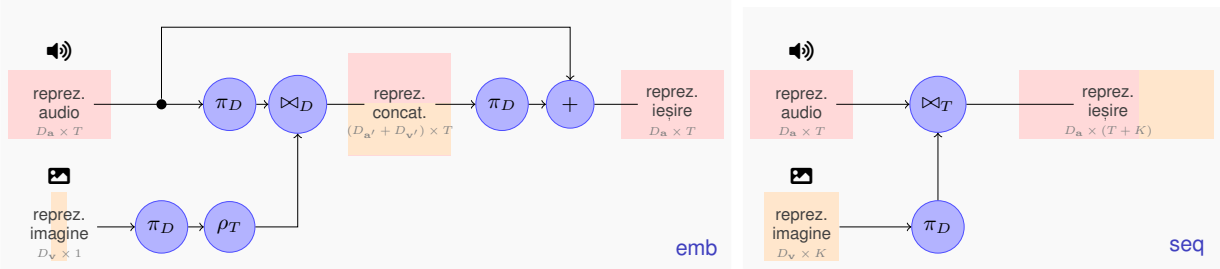


Figura 2: Cele două mecanisme propuse pentru fuziunea modalităților audio și vizuale: emb, fuziune de-a lungul dimensiunii de reprezentării (stânga); seq, fuziune de-a lungul dimensiunii secvenței (dreapta). Operațiile suplimentare (proiecție densă, notată cu π; operațiune de repetare, notată cu ρ) asigură potrivirea dimensiunilor și reprezentări mai bine adaptate; indicele (D sau T) indică axa de-a lungul căreia se aplică fiecare transformare (dimensiunea reprezentării sau dimensiunea secvenței). Simbolul "$\bowtie$" denotă concatenarea.

dacă folosim o listă de caracteristici vizuale ($K > 1$), atunci concatenarea de-a lungul dimensiunii secvenței este mai naturală.

Fuziunea de-a lungul dimensiunii reprezentării (emb). În acest caz, combinăm trăsăturile de vorbire și vizuale de-a lungul dimensiunii reprezentării. Mai precis, proiectăm mai întâi cele două intrări în noi subspații (de dimensiune $D_{\mathbf{a}'}$, respectiv $D_{\mathbf{v}'}$), replicăm reprezentarea vizuală de T ori (de-a lungul axei temporale), concatenăm cele două reprezentări și, în final, proiectăm rezultatul să aibă dimensiunea $D_{\mathbf{a}}$. În acest caz, procedura de fuziune produce o matrice de aceeași dimensiune cu matricea de intrare, $D_{\mathbf{a}} \times T$. Păstrarea dimensiunii inițiale prezintă o serie de avantaje: ne permite să menținem aceeași dimensiune a decodorului ca în cazul unimodal (permițând învățarea prin transfer și comparații mai corecte) și să folosească conexiuni reziduale (de la reprezentările vorbirii la caracteristicile fuzionate, vorbire–imagine), conexiuni care facilitează procesul de antrenare.

Fuziunea de-a lungul dimensiunii secvenței (seq). Atunci când reprezentările celor două modalități de intrare sunt ambele secvențe, o abordare posibilă este de a concatena caracteristicile vizuale și de vorbire de-a lungul dimensiunii secvenței (temporală pentru vorbire și *patch-wise* pentru imagine). Deoarece decodorul "citește" reprezentările de-a lungul dimensiunii secvenţiale a intrării, această operație face ca acest tip de fuziune să fie mai costisitor. Cu toate acestea, fuziunea seq are avantajul de a fi mai flexibilă decât varianta emb, întrucât decodorul poate agrega separat reprezentările audio și vizuale, fără a le amesteca.

Relația cu lucrările anterioare. Multe dintre abordările anterioare s-au bazat pe rețele recurente iar metodele cele mai comune de încorporare a contextului vizual au fost *(i)* setarea primul „cuvânt" decodat cu reprezentarea vizuale [18, 38] sau *(ii)* inițializarea stării ascunsă a straturilor recurente cu reprezentarea vizuale [1]. O altă metodă, întâlnită în special pentru adaptarea caracteristicilor acustice, a fost antrenamentul adaptiv vizual [17, 25], ceea ce echivalează cu aplicarea unei transformări liniare parametrizată de reprezentarea vizuală. În timp ce concatenarea caracteristicilor a fost folosită anterior [17, 25], aceasta nu a fost folosită în contextul arhitecturilor de tip Transformer. Când informația vizuală este o secvență, metodele bazate pe atenție sunt o alegere populară [4, 27, 37]. Pentru toate aceste metode decodorul agregă în mod independent fluxurile audio și vizuale, în timp ce în cazul nostru metoda seq agregă ambele modalități simultan. Metodele din [4, 37] agregă secvența vizuală pe baza cuvântului decodat anterior (cum facem noi), în timp ce [27] se bazează pe sunet pentru operația de agregare. Principala distincție între [4] și [37] este că prima concatenează pur și simplu cele două reprezentări agregate, în timp ce acesta din urmă prezice care dintre cele două modalități (vizuală sau audio) ar trebui să fie preferată printr-un al doilea strat de atenție, ierarhic.

3.2 Setup experimental

Descriem bazele de date utilizate și oferim mai multe detalii privind implementarea.

Baze de date. Deoarece scopul proiectului îl constituie scenariul multimodal, am evaluat sistemul de RAV preponderent baze de date multimodale, care pe lângă audio și transcrierile aferente, asociază și o componentă vizuală—imagini sau video. Bazele de date utilizate sunt următoarele:

- **LibriSpeech** [26] este un corpus de aproximativ 1,000 de ore de cărți audio citite derivate din proiectul LibriVox. Am folosit acest set de date pentru preantrenarea modelului de RAV. Mai exact, am antrenat sistemul

Tabelul 1: Comparație cu abordările din starea artei pe seturile de testare a trei baze de date multimodale (Flickr8K, How2 și Localized Narratives) în funcție de rata de eroare la nivel de cuvânt (valorile mai mici sunt mai bune). Coloana "vizual" indică acele variante care folosesc fluxul vizual ca intrare adițională față de fluxul de vorbire.

metodă	vizual	fuziune	Flickr8K	How2	Loc. Nar.
[38]			14.8	—	—
	✔		13.8	—	—
[37]			13.6	—	—
	✔		14.1	—	—
[27]			—	19.2	—
	✔		—	18.4	—
[4]			—	17.7	—
	✔		—	17.2	—
pretrain			11.1	26.9	49.3
finetune			**3.8**	11.8	4.3
finetune	✔	emb	4.3	11.1	**3.9**
finetune	✔	seq	4.7	**10.8**	4.0

pe cele trei părți standard: `clean100`, `clean360` și `other500`. Eroarea la nivel de cuvânt a modelului preantrenat pe părțile de evaluare `clean` și `other` este de 2.6%, respectiv 6.0%.

- **Flickr8K** [6, 8] conține 8,000 de imagini, fiecare descrisă de câte 5 propoziții (deci în total 40,000 de propoziții). Setul de date original [8] conținea doar imagini și descrierea textuală, dar a fost ulterior extins cu înregistrări audio ale acestor descrieri de către Harwath and Glass [6].
- **How2** [34] conține video-uri de instrucție și învățare descărcate de pe YouTube și vine cu informații suplimentare privitor la segmente temporale (shot-uri video) și transcrieri. Folosim varianta de 300 de ore, care totalizează aproximativ 13,5K videoclipuri (respectiv 190K segmente). Setul de date constă din trăsături audio și vizuale pre-extrase, dar, pentru a putea folosi modelul preantrenate, a trebuit să folosim videoclipurile originale; datele originale ne-au fost furnizate de către autori la cerere.
- **Localized narratives** [29] este un set de date introdus recent, care extinde patru seturi de date de imagini populare (Flickr30K [42], COCO [15], ADE20K [44], Open Images [14]) cu noi descrieri, înregistrări audio și urme de mouse (care localizează cuvintele rostite în imagine). Comparativ cu seturile de date originale, subtitrările sunt mai bogate, iar componenta audio vine cu noi provocări din cauza condițiilor de înregistrare zgomotoase și a vorbirii cu accent. Folosim acest set de date pentru a efectua un studiu de ablație peste configurațiile rețelei de procesare audio. Pentru a facilita efectuarea de studii ample, efectuăm următoarele pre-procesări: *(i)* utilizăm numai componenta Flickr30K, *(ii)* segmentăm audio-ul în propoziții (pe baza transcrierilor furnizate), *(iii)* eliminăm înregistrările mai lungi de 15 secunde, *(iv)* utilizăm jumătate din înregistrări alese aleator. Această procedură se aplică pe toate cele trei set-uri (antrenare, validare, testare), producând aproximativ 32K, 1K, respectiv 1K mostre.

Detalii de implementare. Implementarea noastră se bazează pe utilitarul ESPnet [40]. După cum am menționat, componenta de recunoaștere a vorbirii este o arhitectură de tip Transformer, pe care am preantrenat-o pe setul de date LibriSpeech [26]. Sistemul emite jetoane dintr-un vocabular cu 5,000 de elemente, care a fost obținut prin segmentarea cuvintelor folosind un model de limbă de tip unigram [13]; menționăm că nu folosim un model de limbă extern. Pentru tehnica de reglaj fin, antrenăm modelul preantrenat pentru încă 50 de epoci pe seturile de date mai mici (Flickr8K și Localized Narratives) și 30 de epoci pentru setul de date mai mare (How2). Pentru optimizare, rata de învățare este încălzită liniar de la 3.2×10^{-8} la 8×10^{-4} peste 25K loturi, după care se micșorează ca o funcție de $1/s^2$ în numărul pasului s. La momentul testării, asamblăm cele mai bune zece modele făcând o medie a ponderilor acestora; această tehnică oferă îmbunătățiri mici dar consistente față de predicția doar cu cel mai bun model.

3.3 Rezultate experimentale

Această parte a raportului prezintă evaluarea empirică a sistemului dezvoltat în cadrul proiectului: în secțiunea 3.3.1 comparăm sistemul multimodal cu metodele din starea artei și variantele de bază (unimodale); în secțiunea 3.3.2 prezentăm un studiu experimental asupra principalelor contribuții individuale: învățarea prin transfer și augmentarea datelor; în secțiunea 3.3.3 includem rezultate concrete calitative; secțiunea 3.3.4 prezintă o discuție mai generală a abordării propuse și posibile limitări ale acesteia.

Tabelul 2: Evaluarea impactului augmentării datelor audio (coloana "aug.") pe seturile de testare ale celor trei baze de date multimodale în funcție de rata de eroare la nivel de cuvânt. Toate modelele folosesc tehnica reglajului fin (finetuned), iar variantele multimodale folosesc codorul vizual ResNet50.

vizual	fuziune	aug.	Flickr8K	How2	Loc. Nar.
	—		**3.8**	11.8	**4.3**
	—	✔	4.2	**11.2**	4.5
✔	emb		4.8	11.8	4.1
✔	emb	✔	**4.3**	**11.1**	**3.9**
✔	seq		**4.0**	11.8	4.2
✔	seq	✔	4.7	**10.8**	**4.0**

3.3.1 Rezultate principale

Tabelul 1 prezintă performanța pentru sarcina de recunoaștere a vorbirii pentru patru dintre sistemele noastre: două variante unimodale (un sistem de tip pretrained, folosit ca metodă de bază, și varianta de tip finetuned, reprezentând reglajul fin al variantei pretrained pentru fiecare set de date) și două variante multimodale (ambele antrenate prin reglarea fină a tuturor componentelor, dar care diferă în tehnicile de fuziune, emb sau seq, așa cum am descris în secțiunea 3.1.3). Ambele metode multimodale folosesc tehnica de augmentare a datelor SpecAugment și un ResNet cu 50 de straturi utilizat pe post de codor vizual. Comparăm metodele noastre cu mai multe metode din stare artei. Studiile anterioare evaluează de obicei pe un singur set de date, de exemplu, Flickr8K [37, 38] sau How2 [4, 27], în timp ce noi raportăm performanța pe trei seturi de date: cele două menționate mai sus și Localized Narratives (pe care suntem primii care raportăm rezultate pentru sarcina de recunoaștere a vorbirii din semnale multimodale).

Observăm că metoda pretrained îmbunătățește rezultatele față de lucrările anterioare pe Flickr8K, deși rezultatele sale sunt mai slabe pe How2 și Localized Narratives din cauza nepotrivirii datelor. Cu toate acestea, prin metoda finetuning, sistemul unimodal (bazat doar pe vorbire) depășește semnificativ stadiul actual al stării artei, generând îmbunătățiri relative de 72% și 31% pe Flickr8K, respectiv, How2. Rezultatele pentru sistemele multimodale, care includ informațiile vizuale, sunt mai bune decât rezultatele unimodale în cazul setului de date How2 și Localized Narratives; pentru Flickr8K este dificil de îmbunătățit performanța de bază, probabil, deoarece este un set de date curat pentru care sistemul de RAV funcționează deja bine și multe dintre erorile sale au bază vizuală. Dintre cele două metode de fuziune rezultatele sunt oarecum neconcludente, fuziunea de-a lungul dimensiunii de încorporare, emb, fiind metoda mai bună pe două dintre cele trei seturi de date.

3.3.2 Experimente sistematice

În această subsecțiune prezentăm două studii sistematice care vizează impactul augmentării datelor și importanța transferului reprezentărilor.

Augmentarea datelor. Evaluăm impactul tehnicii de augmentării a datelor în trei scenarii: pentru sistemul unimodal și pentru cele două variante multimodale, care folosesc cele două tehnici de fuziune a caracteristicilor (emb și seq). Pentru toate cazurile, efectuăm reglajul fin al tuturor componentelor, iar pentru sistemele bazate pe fluxul vizual folosim rețeaua ResNet cu 50 de straturi.

Tabelul 2 arată că augmentarea datelor de vorbire este importantă pentru sistemele multimodale, producând îmbunătățiri în cinci din cele șase cazuri. Aceste rezultate sugerează că perturbarea semnalului audio ajută modelele multimodale să se bazeze mai mult pe fluxul vizual și, în cele din urmă, să producă rezultate mai bune. În mod surprinzător, pentru variantele unimodale am observat doar îmbunătățiri limitate prin augmentarea datelor, cu excepția rezultatelor setului de date How2.

Transferul reprezentărilor. În continuare, prezentăm un studiu amplu asupra impactului inițializării (aleatorie sau preantrenată) și procedura de antrenament (ponderi fixe sau ajustate) pentru fiecare dintre componentele modelului (codor audio, codor vizual, decodor). Aceste experimente sunt efectuate pe setul de date Localized Narratives iar pentru setarea multimodală folosim doar metoda de fuziune emb. În plus, investigăm impactul capacității codorului vizual prin varierea numărului de straturi din arhitectura ResNet: 18 sau 50. Rezultatele sunt prezentate în Tabelul 3.

Rândurile 1–5 arată rezultatele pentru sistemul unimodal, corespunzător unui sistem de RAV standard, bazat doar pe vorbire. Observăm că varianta pretrained, fără nicio reglare fină (rândul 1) este slab performantă, cel mai probabil, din

Tabelul 3: Transferarea reprezentărilor—evaluare pe setul de testare al bazei de date Localized Narratives. Pentru fiecare dintre cele trei componente ale modelului (codor audio, codor vizual, decodor), indicăm modul în care ponderile modelului sunt inițializate (fie aleatoriu, fie transferate de la un model preantrenat) și antrenate (fie fixe, fie ajustate prin tehnica reglajului fin cu). Pentru fiecare setare raportăm rata de eroare la nivel de cuvânt (eng., *word error rate*, WER) și numărul de parametri antrenabili. Informația vizuală este fuzionată cu metoda emb. Pentru aceste experimente, nu am folosit tehnica de augmentarea a datelor audio.

	codor audio		codor vizual			decodor		WER (%)	num. parameteri antrenabili ($\times 10^6$)
	init	antrenare	init	antrenare	rețea	init	antrenare		
1	↷	🔒	—	—	—	↷	🔒	49.3	0
2	◆	🔓	—	—	—	◆	🔓	22.5	99.4
3	↷	🔒	—	—	—	◆	🔓	9.1	32.9
4	↷	🔒	—	—	—	↷	🔓	6.3	32.9
5	↷	🔓	—	—	—	↷	🔓	4.3	99.4
6	↷	🔒	↷	🔒	ResNet18	↷	🔓	5.8	33.2
7	↷	🔒	↷	🔓	ResNet18	↷	🔓	5.5	44.4
8	↷	🔓	↷	🔒	ResNet18	↷	🔓	4.3	99.6
9	↷	🔓	↷	🔓	ResNet18	↷	🔓	4.2	110.8
10	↷	🔒	↷	🔒	ResNet50	↷	🔓	5.9	33.4
11	↷	🔒	↷	🔓	ResNet50	↷	🔓	5.6	56.9
12	↷	🔓	↷	🔒	ResNet50	↷	🔓	4.1	99.8
13	↷	🔓	↷	🔓	ResNet50	↷	🔓	4.1	123.3

cauza nepotrivirii mari atât în ceea ce privește datele audio (vorbirea este zgomotoasă și cu accent) cât și din cauza discrepanțelor la nivelul modelului de limbă. Pe de altă parte, ignorarea disponibilității reprezentărilor preantrenate (rândul 2) nu este, de asemenea, ideală: antrenarea rețelei de la zero, așa cum se face de obicei în lucrările anterioare, produce transcrieri mai bune, dar încă nesatisfăcătoare. Rândurile 3 și 4 arată rezultatele pentru cazul în care codorul este fix și decodorul este antrenat: fie de la zero (rândul 3), fie prin reglarea fină a greutăților preantrenate (rândul 4). Deoarece decodorul unui model RAV de tip end-to-end joacă și rolul unui model de limbă, această procedură este asemănătoare cu adaptarea limbii și are ca rezultat creșterea semnificativă a performanței pentru ambele variante. În cele din urmă, reglarea fină a ambelor componente (rândul 5) dă cele mai bune rezultate, cu o îmbunătățire relativă de aproximativ 30%.

Rândurile 6–13 prezintă rezultatele pentru sistemele multimodale, folosind un codor vizual cu 18 (rândurile 6–9), respectiv 50 de straturi (rândurile 10–13). Pentru acest set de experimente, folosim doar cu abordarea pe bază de reglaj fin, deoarece rezultatele sistemului unimodal au arătat că această tehnică este superioară. De asemenea, adaptăm întotdeauna decodorul deoarece fuziunea vorbire-viziune afectează distribuția statistică a reprezentărilor. Este de reținut faptul că reprezentările fuzionate sunt proiectate la aceeași dimensiune de încorporare ca și caracteristicile de vorbire, ceea ce permite transferarea ponderilor decodorului pe baza modelului preantrenat. Straturile de proiecție din stratul de fuziune sunt întotdeauna antrenabile.

Mai întâi observăm că includerea informațiilor vizuale îmbunătățește performanța față de sistemul cu flux unic în toate scenariile: fie dacă menținem codoarele fixe (rândurile 6 și 10 față de rândul 4), fie dacă aplicăm tehnica de reglaj fin asupra codoarelor (rândurile 9 și 13 față de rândul 5). În al doilea rând, observăm că obținem rezultate mai bune pe măsură ce permitem ajustarea mai multor componente, ultima coloană indicând o corelație între numărul de parametri antrenabili și performanță. Cele mai bune rezultate se obțin la reglarea fină a tuturor componentelor (rândurile 9 și 13). De asemenea, creșterea capacității codorului vizual rezultă în concluzii similare. Observăm ușoare îmbunătățiri pentru cazurile în care reglam cu precizie codorul asociat vorbirii (rândul 12 față de rândul 8; rândul 13 față de rândul 9), potențial sugerând cuplarea dintre cele două modalități trebuie să fie luată în considerare și de codor și nu doar de decodor.

3.3.3 Rezultate calitative

În această subsecțiune prezentăm o serie de exemple calitative pentru a compara sistemul unimodal cu cel multimodal. Rezultatele sunt furnizate pe setul de date How2 folosind modele antrenate în regimul finetune și folosind augmentarea datelor; pentru modelul multimodal am folosit metoda de fuziune emb și codorul vizual ResNet50. Tabelul 3 prezintă

r · mix it up really good because that egg white is thick it's really thick

u · mix it up really good because that *eight* white is thick it's really thick

m · mix it up really good because that egg white is thick it's really thick

r · and spalting is nothing more than the natural decay process that wood goes through

u · and spalting is nothing more than the natural decay process that *would* goes through

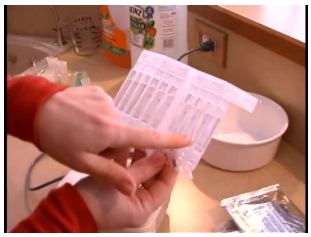
m · and spalting is nothing more than the natural decay process that wood goes through

r · so we can either take the fiber wire or use the shotgun or use the pistol that we picked up

u · so we can either take the fiber wire or use the *shock on* or use the pistol that we picked up

m · so we can either take the fiber wire or use the shotgun or use the pistol that we picked up

---



r · so each vial here is actually one use

u · so each *vile* here is actually one use

m · so each vial here is actually one use

r · that's just a long road

u · that is just a long *row*

m · that's just a long road

r · never leave your person on their back

u · never leave your person *on your* on their back

m · never leave your person on their back

r · here's an example of a nicely bound script

u · here's an example of a nicely bound script

m · here's an example of a nicely *balance* script

r · i just took our salad out of the refrigerator

u · i just took our salad out of the refrigerator

m · *now* i just took our salad out of the refrigerator

r · five more keep breathing deep expanding

u · five more keep breathing deep expanding

m · five more *key breathe* than deep expanding

Figura 3: Rezultate calitative pe setul de date How2. Pentru fiecare exemplu, arătăm cadrul central al filmării video, textul de referință (r), transcripțiile folosind modelul unimodal (u) și cel multimodal (m). Greșelile sunt afișate cu roșu și cursiv. Modelul multimodal folosește setarea finetune, metoda de fuziune emb și codificatorul vizual ResNet50.

nouă mostre în care cele două sisteme diferă în predicțiile lor, dar cel puțin unul este corect: pentru primele două rânduri sistemul multimodal este corect, în timp ce în ultimul rând sistemul unimodal este corect. Primele patru exemple sunt cazuri în care sistemul multimodal alege corect între cuvinte similare fonetic, de exemplu, corectează "eight" cu "egg", "would" cu "wood", "shock on" cu "shotgun", "vile" cu "vial". În timp ce aceste exemple sugerează că contextul vizual ajută la transcriere, restul mostrelor sunt mai greu de interpretat.

3.3.4 Analiză și discuție

În continuare discutăm două potențiale limitări ale abordării noastre. Aceste neajunsuri provin din natura problemei, care implică o modalitate (fluxul audio) care o domină pe cealaltă (fluxul vizual).

Îmbunătățiri limitate. Deși studiul experimental arată că rezultatele tind să se îmbunătățească în cadrul multimodal, dar o posibilă critică la adresa abordării noastre ar putea fi că creșterea în performanță nu este semnificativă. Cu toate acestea, rezultatele noastre sunt în conformitate cu cele raportate de lucrările din starea artei (vezi Tabelul 1), care funcționează într-o configurație mai favorabilă, deoarece variantele lor multimodale au mai multe posibilități de îmbunătățire (deoarece pornesc de la sisteme de bază inferioare). Mai mult, credem că este dificil să obținem îmbunătățiri considerabil mai mari în cadrul actual. Deși cuantificarea cu precizie a limitei superioare a performanței multimodale este o sarcină dificilă, am inspectat manual cele mai frecvente greșeli care apar în cadrul unimodal (luăm în considerare setul de date How2 și metoda unimodală antrenată cu opțiunea finetune și folosind augmentarea datelor). Observăm că multe dintre cele mai frecvente greșeli implică cuvinte scurte care corespund prepozițiilor și pronumelor (cum ar fi "and", "to", "you", "I", "the", "a", "that"), care sunt fie eronat introduse sau omise în transcrierea automată. Aceste tipuri de greșeli sunt greu de corectat prin modalitatea vizuală, deoarece aceste cuvinte nu au o bază vizuală. Mai puțin frecvente sunt greșeli care au context vizual, cum ar fi perechile prezentate în rezultatele calitative sau greșeli de substituție precum "cymbal" $\rightarrow$ "symbol", "sprite" $\rightarrow$ "sprayed", "bow" $\rightarrow$ "boat" sau "both". Cu toate acestea, acest tip de erori sunt în minoritare, ceea ce indică faptul că îmbunătățiri substanțiale sunt dificil de realizat.

Rolul modalității vizuale. Rezultatele calitative indică faptul că există cazuri pentru care rămâne neclar *cum* sistemul multimodal folosește componenta vizuală. Lucrările anterioare în contextul recunoașterii vorbirii multimodale [35] și al traducerii automate multimodale [41] au observat că fluxul vizual ajută în moduri neașteptate. În special, Wu *et al.* [41] sugerează că partea vizuală joacă rolul unui regularizator și nu injectează neapărat informații utile în sistem. În cazul nostru, deși credem că creșterea propusă a datelor de vorbire poate atenua problema într-o oarecare măsură, efectul său poate rămâne încă insuficient. Presupunem că problema constă în cuplarea insuficient de strânsă a modalităților de intrare. O posibilă soluție pentru o fuziune mai pervazivă ar fi pregătirea preliminară a unui model audio-vizual auto-supravegheat pe cantități mari de date. Astfel de sisteme audio-vizuale au devenit comune, dar ele nu au fost aplicate în setarea de recunoaștere multimodală a vorbirii. Cele mai apropiate lucrări în această direcție sunt cele ale lui Hsu *et al.* [9], care utilizează reprezentările audio-vizuale pregătite în prealabil pentru recunoașterea unimodală a vorbirii și ale lui Rouditchenko *et al.* [31], care efectuează sarcina de regăsire (eng., *retrieval*) multimodală a textului.

4 Impactul proiectului

În acest raport, am propus și dezvoltat metode pentru recunoașterea multimodală a vorbirii. Arhitectura sistemului o constituie un sistem de recunoaștere a vorbirii de tip Transformer în care am injectat informații vizuale printr-un codor de imagine de tip ResNet. Spre deosebire de metodele anterioare, am utilizat reprezentări preantrenate atât pentru codorul audio, cât și pentru cel vizual, și am explorat două tehnici de fuziune ale informațiilor produse de cele două codoare. Cel mai important rezultat l-au reprezentat îmbunătățirile substanțiale obținute față de starea artei pe două seturi de date multimodale standard: Flickr8K și How2. De asemenea, studiile sistematice efectuate asupra fiecăreia dintre componente indică importanța alegerilor noastre: datele audio augmentate pentru antrenarea rețelei multimodale şi antrenarea de tip end-to-end (prin tehnica de reglaj fin a tuturor componentelor ale sistemului). Deși considerăm remarcabil faptul că configurația multimodală îmbunătățește rezultatele față de sistemul unimodal deja foarte performant, exemplele calitative lasă încă deschisă întrebarea cum exact folosește sistemul informațiile vizuale.

Direcții viitoare. Pentru a înțelege mai bine interacțiunile complicate care apar într-o rețea multimodală de recunoaștere a vorbirii, o direcție viitoare de studiu ar putea investiga care părți ale intrărilor (audio, imagine, simboluri anticipate anterior) contribuie mai mult la ieșire; unelte moderne pentru generarea de explicații în învățarea automată [10, 12, 33] reprezintă o posibilă abordare. O altă direcție viitoare ar putea implica cuplarea celor două codoare printr-o fuziune mai pervazivă prin valorificarea modelelor audio-vizuale auto-supravegheate care sunt antrenate în prealabil pe seturi de date la scară largă.

Potențiale aspecte sociale negative. Scopul acestei cercetări este de a face recunoașterea vorbirii mai precisă atunci când sunt disponibile informații vizuale suplimentare. Recunoașterea vorbirii este o tehnologie care permite multe aplicații de impact, cum ar fi generarea automată de subtitrări sau căutarea de cuvinte cheie în colecții mari de date audio. Deși, ca orice abordare a învățării automate, aceasta poate fi supusă unor prejudecăților ce apar sistematic în date (eng., *data bias*) sau utilizării rău intenționate, nu prevedem niciun aspect social negativ care ar putea decurge specific din direcția actuală a proiectului.

Director proiect,
Dan Oneață

Bibliografie

[1] O. Caglayan, R. Sanabria, S. Palaskar, L. Barraul, and F. Metze. Multimodal grounding for sequence-to-sequence speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8648–8652, 2019.

[2] A. Caranica, D. Oneață, H. Cucu, and C. Burileanu. Confidence estimation for lattice-based and lattice-free automatic speech recognition. *UPB Scientific Bulletin, Series C: Electrical Engineering and Computer Science*, 83(3):155–170, 2021.

[3] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[4] S. Ghorbani, Y. Gaur, Y. Shi, and J. Li. Listen, look and deliberate: Visual context-aware speech recognition using pre-trained text-video representations. In *IEEE Spoken Language Technology Workshop*, pages 621–628, 2021.

[5] A. Gupta, Y. Miao, L. Neves, and F. Metze. Visual features for context-aware speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5020–5024, 2017.

[6] D. Harwath and J. Glass. Deep multimodal semantic embeddings for speech and images. In *Workshop on Automatic Speech Recognition and Understanding*, pages 237–244, 2015.

[7] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[8] M. Hodosh, P. Young, and J. Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899, 2013.

[9] W.-N. Hsu, D. Harwath, and J. Glass. Transfer learning from audio-visual grounding to speech recognition. In *Interspeech*, pages 3242–3246, 2019.

[10] G. Joshi, R. Walambe, and K. Kotecha. A review on explainability in multimodal deep neural nets. *IEEE Access*, 2021.

[11] S. Karita, N. Chen, T. Hayashi, T. Hori, H. Inaguma, Z. Jiang, M. Someki, N. E. Y. Soplin, R. Yamamoto, X. Wang, S. Watanabe, T. Yoshimura, and W. Zhang. A comparative study on transformer vs RNN in speech applications. In *Workshop on Automatic Speech Recognition and Understanding*, pages 449–456, 2019.

[12] N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Melnikov, N. Kliushkina, C. Araya, S. Yan, et al. Captum: A unified and generic model interpretability library for PyTorch. *arXiv preprint arXiv:2009.07896*, 2020.

[13] T. Kudo. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Association for Computational Linguistics*, pages 66–75, 2018.

[14] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, A. Kolesnikov, et al. The Open Images dataset v4. *International Journal of Computer Vision*, pages 1–26, 2020.

[15] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[16] H. Liu, Z. Dai, D. R. So, and Q. V. Le. Pay attention to MLPs. *arXiv preprint arXiv:2105.08050*, 2021.

[17] Y. Miao and F. Metze. Open-domain audio-visual speech recognition: A deep learning approach. In *Interspeech*, pages 3414–3418, 2016.

[18] Y. Moriya and G. J. Jones. LSTM language model adaptation with images and titles for multimedia automatic speech recognition. In *IEEE Spoken Language Technology Workshop*, pages 219–226, 2018.

[19] Y. Moriya and G. J. Jones. Multimodal speaker adaptation of acoustic model and language model for ASR using speaker face embedding. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8643–8647, 2019.

[20] K. Olaleye, D. Oneață, and H. Kamper. Keyword localisation in untranscribed speech using visually grounded speech models. *IEEE Journal of Selected Topics in Signal Processing*, 2022.

[21] D. Oneață and H. Cucu. Multimodal speech recognition for unmanned aerial vehicles. *Computers & Electrical Engineering*, 90:106943, 2021.

[22] D. Oneață and H. Cucu. Improving multimodal speech recognition by data augmentation and speech representations. In *Computer Vision and Pattern Recognition Workshops*. IEEE, 2022.

[23] D. Oneață, A. Stan, and H. Cucu. Speaker disentanglement in video-to-speech conversion. In *29th European Signal Processing Conference*, 2021.

[24] D. Oneață, B. Lőrincz, A. Stan, and H. Cucu. FlexLip: A controllable text-to-lip system. *Sensors*, 22(11), Jun 2022.

[25] S. Palaskar, R. Sanabria, and F. Metze. End-to-end multimodal speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5774–5778, 2018.

[26] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur. Librispeech: An ASR corpus based on public domain audio books. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5206–5210, April 2015.

[27] G. Paraskevopoulos, S. Parthasarathy, A. Khare, and S. Sundaram. Multiresolution and multimodal speech recognition with transformers. In *Association for Computational Linguistics*, 2020.

[28] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le. SpecAugment: A simple data augmentation method for automatic speech recognition. In *Interspeech*, pages 2613–2617, 2019.

[29] J. Pont-Tuset, J. Uijlings, S. Changpinyo, R. Soricut, and V. Ferrari. Connecting vision and language with localized narratives. In *European Conference on Computer Vision*, pages 647–664, 2020.

[30] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

[31] A. Rouditchenko, A. Boggust, D. Harwath, D. Joshi, S. Thomas, K. Audhkhasi, R. Feris, B. Kingsbury, M. Picheny, A. Torralba, et al. AVLnet: Learning audio-visual language representations from instructional videos. In *Interspeech*, 2021.

[32] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 12 2015. ISSN 0920-5691. doi: 10.1007/s11263-015-0816-y.

[33] W. Samek, G. Montavon, S. Lapuschkin, C. J. Anders, and K.-R. Müller. Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE*, 109(3):247–278, 2021.

[34] R. Sanabria, O. Caglayan, S. Palaskar, D. Elliott, L. Barrault, L. Specia, and F. Metze. How2: A large-scale dataset for multimodal language understanding. In *Advances in Neural Information Processing Systems*, 2018.

[35] T. Srinivasan, R. Sanabria, and F. Metze. Analyzing utility of visual context in multimodal speech recognition under noisy conditions. In *The How2 Challenge: New Tasks for Vision & Language, ICML*, 2019.

[36] T. Srinivasan, R. Sanabria, and F. Metze. Looking enhances listening: Recovering missing speech using images. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6304–6308, 2020.

[37] T. Srinivasan, R. Sanabria, F. Metze, and D. Elliott. Fine-grained grounding for multimodal speech recognition. In *Empirical Methods in Natural Language Processing*, 2020.

[38] F. Sun, D. Harwath, and J. Glass. Look, listen, and decode: Multimodal speech recognition with images. In *IEEE Spoken Language Technology Workshop*, pages 573–578, 2016.

[39] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

[40] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. E. Y. Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai. ESPnet: End-to-end speech processing toolkit. In *Interspeech*, pages 2207–2211, 2018.

[41] Z. Wu, L. Kong, W. Bi, X. Li, and B. Kao. Good for misconceived reasons: An empirical revisiting on the need for visual context in multimodal machine translation. In *Association for Computational Linguistics*, 2021.

[42] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

[43] A. Zeyer, P. Bahar, K. Irie, R. Schlüter, and H. Ney. A comparison of Transformer and LSTM encoder decoder models for ASR. In *Workshop on Automatic Speech Recognition and Understanding*, pages 8–15, 2019.

[44] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, and A. Torralba. Semantic understanding of scenes through the ADE20K dataset. *International Journal of Computer Vision*, 127(3):302–321, 2019.