Vorbis: Recunoașterea automată a vorbirii din semnale multi-modale

Raport tehnico-științific · etapa 2 / 2021

Dan Oneață
Universitatea POLITEHNICA din București
dan.oneata@gmail.com

1 Rezumatul etapei

Conform planului de realizare pentru etapa 2 / 2021 am avut de livrat următoarele rezultate:

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.

Aceste obiective au fost atinse în proporție de 100% și au fost realizate prin cele trei activități definite în planul de realizare:

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 [14, 42, 46] și implementat cu ajutorul utilitarului ESPnet [44]. Am adus îmbunătățiri sistemului de bază utilizând tehnici precum învățarea prin transfer (pornind de la baza de date LibriSpeech [23]) și augmentarea de date [25]. 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 [31] atingem o eroare la nivel de cuvânt de 11.2%, față de 17.7% cât a fost raportat anterior de Ghorbani et al. [9]).

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 [11] pre-antrenată pentru clasificare pe baza de date ImageNet [30] și *(ii)* arhitecturile ResNet și Visual Transformer [7] 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 [28]. 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 [12]), 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. Toate aceste experimente au avut ca scop evaluarea capacității de înțelegere și captare a informației vizuale, urmând ca în cursul etapei următoare să integrăm arhitecturile vizuale prin diverse tehnici de fuziune în sarcina de interes—recunoașterea automată a vorbirii din semnale multi-modale.

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

2 Descrierea științifică și tehnică

În continuare prezentăm în detaliu aspectele științifice și tehnice ce corespund celor trei activități desfășurate în cadrul etapei 2 / 2021. Fiecare activitate este elaborată în într-o secțiune dedicată: A2.1 în secțiunea 2.1, A2.2 în secțiunea 2.2, A2.3 în secțiunea 2.3.

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

În această secțiune prezentăm sistemul de recunoaștere automată a vorbirii (RAV) de tip end-to-end pe care l-am dezvoltat în cadrul proiectului. Scopul sarcinii de RAV este de a produce în mod automat transcrierea (cuvintele pronunțate) pentru o înregistrare audio a unei uterănțe. Arhitectura pe care o propunem este de tip Transformer [42], care este deja consacrată în procesarea de limbaj natural [6, 27] și procesarea de vorbire [14, 46]. Acest tip de arhitectură are două componente de bază: un codor, care procesează fluxul audio, și un decodor, care produce transcrierea pe baza informațiilor oferite de codor. Codorul se bazează pe module de auto-atenție (eng., *self-attention*) pentru a învăța reprezentări alte semnalului audio, iar decodorul se bazează pe module de atenție încrucișată (eng., *cross-attention*) pentru a incorpora informațiile din fluxul audio. Rețeaua prezice transcrierea într-o manieră autoregresivă prin modelarea probabilității $p(t_k|t_{1:k}, \mathbf{a})$ a următoarei "bucăți" de text t_k dat audio-ul $\mathbf{a}$ și transcrierea parțială anterioară $t_{1:k}$. Transcrierea se face la nivel de jetoane (eng., *tokens*), care reprezintă unități mai mici decât cuvintele (eng., *subwords*), deci textul produs la ieșire este sub forma unei secvențe de jetoane $t_1, \ldots, t_n$, numărul acestora n variind în funcție de lungimea uterănței rostite. Utilizarea de jetoane în loc de cuvinte oferă flexibilitate, deoarece sistemul poate să producă cuvinte noi, care nu sunt predefinite într-un vocabular fix. Pentru a învăța parametri arhitecturii folosim ca funcție de pierdere (eng., *loss function*) valoarea negativă a verosimilității (eng., *negative log likelihood*) peste jetoanele prezise. Menționăm că arhitectura propusă este tip end-to-end pentru că primește la intrare pe semnale audio și scoate la ieșire transcrieri într-un singur pas, fără a trece prin componente intermediare care să necesite antrenament separat, cum este cazul metodelor de RAV de tip hidden Markov model (HMM).

În continuare prezentăm două îmbunătățiri pe care le aducem: învățarea prin transfer și augmentarea datelor audio.

Învățarea prin transfer. În loc de a porni învățarea sistemului de RAV de la zero, explorăm inițializarea celor componente (codor și decodor) pe baza unui model antrenat în prealabil (pre-antrenat). Crucial, pre-antrenarea sistemului de RAV utilizează un set mare cu date de vorbire, în cazul nostru corpus-ul LibriSpeech [23]. Ponderile astfel învățate sunt apoi utilizate ca punct de pornire pentru continuarea antrenării pe setul de date țintă; această procedură poarte numele de tehnica de reglaj fin (eng., *fine-tuning*).

Augmentări de date. Pentru creșterea robusteții sistemului, extindem setul de înregistrări audio cu versiuni perturbate ale acestora. Augmentările se bazează pe tehnica SpecAugment [25] și includ proceduri precum modificarea vitezei vorbirii, mascarea în frecvență și mascarea în timp. Aceleași transformări au fost folosite pentru antrenarea sistemului de bază pe LibriSpeech și le aplicăm și în etapa de reglaj fin. De notat, că această abordare poate fi utilă și în etapa viitoare a proiectului vom construi un sistem de RAV de tip multi-modal, care utilizează la intrare atât audio cât și imagini. Abordările anterioare pentru RAV multi-modal [34, 35, 37] au folosit mascarea în timp, dar în cazul lor segmentele mascate corespundeau anumitor cuvinte (cum ar fi substantive și complemente). Ca atare, aceste metode necesită componente suplimentare, cum ar fi utilitare care să alinieze audio-ul cu transcrierea și care să identifice anumite părți de vorbire. Pe de altă parte, abordarea noastră este mult mai flexibilă, deoarece augmentările propuse sunt lipsite de aceste dependențe. O altă diferență față de abordările anterioare este că acestea au investigat mascarea cu un alt scop (nu ca tehnică de creștere a datelor), ci pentru a cuantifica cât de bine poate componenta vizuală să recupereze cuvintele mascate.

2.1.1 Setup experimental

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

Baze de date. Deoarece scopul proiectului îl constituie scenariul multi-modal, am evaluat sistemul de RAV preponderent baze de date multi-modale, care pe lângă audio și transcrierile aferente, asociază și o componentă vizuală—imagini sau video. Pentru experimentele curente utilizăm doar partea de audio și text din aceste baze de date, dar partea vizuală ne va facilitate extinderea experimentelor pentru activitatea etapei următoare. Bazele de date utilizate sunt următoarele:

- **LibriSpeech** [23] este un corpus de aproximativ 1,000 de ore de cărți audio citite derivate din proiectul LibriVox. Am folosit acest set de date pentru pre-antrenarea modelului de RAV. Mai exact, am antrenat

Tabelul 1: Comparație a sistemului propus de recunoaștere automată a vorbirii față de alte abordări din starea artei pe trei seturi de date: Flickr8K, How2 și Localized Narratives. Raportăm rata de eroare a cuvintelor, deci valorile mai mici sunt mai bune.

metodă	Flickr8K	How2	Loc. Nar.
Sun et al. [38]	14.8	—	—
Srinivasan et al. [36]	13.6	—	—
Paraskevopoulos et al. [24]	—	19.2	—
Ghorbani et al. [9]	—	17.7	—
pre-antrenare	11.1	26.9	49.3
reglaj fin	**3.8**	11.8	**4.3**
reglaj fin + augmentare	4.2	**11.2**	4.5

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

- **Flickr8K** [10, 12] conține 8,000 de imagini, fiecare descrisă de câte 5 propoziții (deci în total 40,000 de propoziții). Setul de date original [12] conținea doar imagini și descrierea textuală, dar a fost ulterior extins cu înregistrări audio ale acestor descrieri de către Harwath and Glass [10].
- **How2** [31] 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 pre-antrenate, a trebuit să folosim videoclipurile originale; datele originale ne-au fost furnizate de către autori la cerere.
- **Localized narratives** [26] este un set de date introdus recent, care extinde patru seturi de date de imagini populare (Flickr30K [45], COCO [18], ADE20K [47], Open Images [17]) 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 [44]. 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 [23]. 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 [16]; menționăm că nu folosim un model de limbă extern. Pentru tehnica de reglaj fin, antrenăm modelul pre-antrenat 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.

2.1.2 Rezultate experimentale

Tabelul 1 prezintă performanța sistemului de recunoaștere automată a vorbirii pentru trei configurații ale sistemelor noastre: un sistem bazat pe un model pre-antrenat pe baza de date LibriSpeech și două variante care implică utilizarea tehnicii de reglaj fin (eng., *fine-tuning*) peste acesta—una dintre variante utilizează date audio augmentate, precum am descris la începutul secțiunii. Contrastăm abordările noastre cu mai multe metode de RAV care au fost folosite ca sisteme de bază în abordările de RAV multi-modal [9, 24, 36, 38]. Studiile anterioare evaluează de obicei pe un singur set de date, de exemplu, Flickr8K [36, 38] sau How2 [9, 24], în timp ce noi raportăm performanța pe trei seturi de date: cele două deja menționate și baza de date Localized Narratives.

Observăm că metoda bazată pe pre-antrenare îmbunătățește deja performanță față de lucrările anterioare pe setul de date Flickr8K, deși rezultatele sale sunt mai slabe pe How2 și Localized Narratives din cauza nepotrivirii datelor. Cu toate acestea, prin tehnica reglajului fin, sistemul de RAV depășește semnificativ starea artei atât pe baza de date

Tabelul 2: Învățarea prin transfer—evaluarăm pe partea de testare a setului de date Localized Narratives. Pentru fiecare dintre cele trei componente ale modelului (codor audio, decodor), indicăm modul în care sunt inițializate ponderile modelului (fie aleatoriu 🔀, fie prin transfer de la un model pre-antrenat ↱) și antrenate (fie sunt fixe 🔒, fie învățate cu 🔓). Pentru fiecare configurație raportăm rata de eroare a cuvintelor și numărul de parametri antrenabili. Pentru aceste experimente, nu am folosit augmentarea datelor audio.

	codor audio		decodor		eroare (%)	număr parametri antrenabili ($\times 10^6$)
	init	train	init	train		
1	↱	🔒	↱	🔒	49.3	0
2	🔀	🔓	🔀	🔓	22.5	99.4
3	↱	🔒	🔀	🔓	9.1	32.9
4	↱	🔒	↱	🔓	6.3	32.9
5	↱	🔓	↱	🔓	4.3	99.4

Flickr8K cât și pe How2. Interesant, metoda bazată pe augmentare produce rezultate contradictorii: pe How2 rezultate se îmbunătățesc (de la 11.8% la 11.2%), dar pe celelalte două seturi de date tendința este inversă. Bănuim că acestea din urmă sunt mai simplificate decât How2, care din cauza dificultății beneficiază de astfel de augmentări.

În tabelul 2 efectuăm un studiu detaliat de ablație asupra impactului inițializării (aleatorie sau preantrenată) și procedura de antrenament (greutăți fixe sau ajustate) pentru fiecare dintre componentele modelului (codor și decodor). Aceste experimente sunt efectuate numai pe setul de date Localized Narratives.

Observăm că varianta care utilizează modelul pre-antrenat, fără tehnica de reglare fină (rândul 1) obține o performanță slabă, cel mai probabil, din cauza nepotrivirii mari atât în ceea ce privește datele audio (vorbirea este zgomotoasă și cu accent), cât și partea lingvistică (modelul de limbă implicit modelului capturând alte statistici). Pe de altă parte, ignorarea disponibilității reprezentărilor pre-antrenate (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 ar ată rezultatele pentru cazul în care codorul este fix și doar decodorul este antrenat: fie de la zero (rândul 3), fie prin reglarea fină a ponderilor pre-antrenate (rândul 4). Deoarece decodorul unui model ASR end-to-end joacă și rolul unui model de limbaj, 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% (față de varianta prezentată pe rândul 4). Dacă analizăm performanța și din punctul de vedere al eficienței (numărul de parametri antrenabili), observăm că varianta în care efectuăm reglaj fin la decodor este preferabilă: obținem cel mai bun compromis eroare–număr de parametri.

2.2 Sistem vizual de tip end-to-end

În această secțiune descriem sistemele vizuale care au fost investigate. Ideea proiectului este de a incorpora informația vizuală în sistemul de recunoaștere de vorbire în speranța îmbunătățirii transcrierii audio-ului de la intrare. În acest scop ne dorim arhitecturi care să fie cât mai flexibile și capabile să învețe reprezentări care sumarizează informația vizuală și care pot fi transferate cu succes la sarcini cât mai diverse. În continuare discutăm arhitecturile propuse, iar în secțiunea experimentală le evaluăm în diferite scenarii. Fezabilitatea arhitecturilor vizuale și transferul reprezentărilor produse de acestea sunt cuantificate în două aplicații: înțelegerea semantică a imaginilor și traducerea video–vorbire. Aceste sarcini sunt multi-modale, dar sunt simplificate față de sarcina de interes RAV multi-modal, care va face obiectul următorului raport.

ResNet (eng., *residual network*) [11] este una din cele mai populare rețele neuronale de clasificare pe imagini și, deși introdusă în 2016, performanța ei este încă la nivelul stării artei [2]. Principala ei inovație o reprezintă conexiunile reziduale—conexiuni de tip scurt-circuit care sar peste un strat, de forma $x + f(x)$—care facilitează antrenarea și optimizarea parametrilor rețelelor neuronale foarte adânci; de exemplu, o versiune a rețelei ajunge la 152 de straturi, de opt ori mai mari decât predecesorul acesteia, rețeaua VGG [33]. În versiunea ei standard, rețeaua ResNet este antrenată pe baza de date ImageNet și prezice una din cele 1,000 de clase de ieșire. Pentru a transfera reprezentările, păstrăm ponderile pre-antrenate pe baza de Imagenet, dar utilizăm activările de straturi anterioare stratului de ieșire. Propunem două variante, în funcție de locația de la care utilizăm aceste activări:

- înainte de ultimul strat (stratul de clasificare), de unde obținem un vector de dimensiune fixă de exemplu 512 pentru ResNet-{18,34} sau 2,048 pentru ResNet-{50,101,152};

- înainte de penultimul strat (stratul de mediere global), de unde obținem o grilă de vector de dimensiune fixă; grila are dimensiune 7×7 (dacă ne asigurăm ca imaginile de intrare au dimensiune 224×224).

Diferența între cele două variante este că cea de-a doua conține informație spațială care credem că poate ajuta modelul să reprezinte mai în detaliu imaginea de intrare.

Aceste reprezentări pot fi în continuare îmbunătățite în funcție de sarcina de interes învățând reprezentări de nivel mai înalt. De exemplu, în cazul în care avem o grilă de vectori putem învăța interacțiuni între locații utilizând o arhitectură de tip perceptron multistrat modulați (eng., *gated multilayer perceptron*; gMLP) [19]. Această rețea este un substitut recent introdus pentru straturile de auto-atenție, care aplică alternativ straturi dense (linear) pe direcția canalelor și cea a secvenței. În comparație cu stratul de atenție, arhitectura gMLP necesită mai puține calcule și memorie, menținând în același timp performanța. Intenționăm să utilizăm acest timp de transformare pentru modelul de recunoaștere automată a vorbirii de tip multi-modal pe care urmează să-l dezvoltăm.

O altă modificare a rețelei ResNet este modificarea straturilor de intrare. În cazul în care la intrare avem video pentru a integra mai bine informația temporală prefixăm un strat de convoluții tridimensionale și sufixăm un strat de tip LSTM pentru a lua din nou în calcul informația temporală anterioară. În secțiunea experimentală evaluăm această arhitectură pentru sarcina de traducere a secvențelor de video mut în vorbire.

CLIP (eng., *contrastive language-image pretraining*) [28] reprezintă un mod de a învăța reprezentări vizuale utilizând supervizia limbajului natural. Concret, pe baza de perechi imagini–descrieri text, se antrenează două codoare—unul vizual pentru imagini, altul pentru limbajul natural—astfel încât să maximizeze acordul între perechile imagini–text aliniate și să minimizeze acordul între perechi alese aleator. O caracteristică importantă a acestei metode este că este antrenată pe un vast set de date, peste 400 milioane de imagini-descrieri culese în mod automat de pe internet. Arhitectura vizuală utilizată de CLIP are două variante: ResNet [11] și Transformer [7]. Autorii au demonstrat capabilități impresionante de transfer, putând generaliza la noi seturi de date fără a fi nevoie de tehnica de reglaj fin. În cazul nostru evaluăm această rețea pentru sarcina de înțelegerea semantică a imaginilor prin găsirea de cuvintelor cheie în imagini descrise.

2.2.1 Sarcina 1: Înțelegerea semantică a imaginilor

Evaluăm arhitecturile vizuale pentru sarcina de înțelegere semantică a imaginilor descrise de oameni. Mai exact, dată o imagine ne dorim să vedem cât de bine identifică o rețea vizuală pre-antrenată cuvinte din descrierea textuală a imaginilor. Ne așteptăm ca performanța pe această sarcină să fie reprezentativă și pentru sarcina de interes (de recunoaștere a vorbirii din semnale multi-modale), deoarece rețeaua vizuală va capta conceptele care apar în imagine și le va pasa mai departe modulului de procesare a semnalului vorbit.

Setul de date. Rulăm evaluarea experimentală pe baza de date Flickr8K [12], pe care am descris-o anterior în subsecțiunea 2.1.1. Fiecare imagine are asociată cinci descrieri textuale pe care le reprezentăm ca un vector binar peste un vocabular de 67 de cuvinte: vectorul conține valoarea 1 pe poziția k dacă descrierea conține al k-lea cuvânt din vocabular; altfel valoarea din vector pe poziția k este 0. Cele 67 de cuvinte au fost alese de Kamper et al. [13] și sunt selectate aleator din cele mai comune 1,000 de cuvinte și sunt adițional verificate că adnotatorii sunt în acord; exemple de cuvinte cheie includ: "*jumps*", "*ocean*", "*race*".

Detalii de implementare. Folosim trei arhitecturi: VGG, ResNet și Visual Transformer (ViT). Utilizăm versiunea VGG-16, aceasta este preantrenată pe ImageNet pentru clasificare a 1,000 de categorii, dar înlocuim stratul de clasificare final cu un perceptron multistrat (MLP) cu 4 straturi de 2,048 de neuroni fiecare și activări de tip ReLU. Această arhitectură prezice peste vocabularul de 67 de cuvinte. La antrenare trunchiul din VGG-16 este ținut fix în tip ce MLP-ul de la final este în continuare prin tehnica de reglaj fin pe bazele de date FlickR30K și Microsoft COCO, care reprezintă alte baze de date conțin perechi imagini–descrieri text. Ultimele două arhitecturi sunt modele pre-antrenate de tip CLIP (descrise anterior); acestea sunt utilizate în scenariul zero-shot, ceea ce înseamnă că modelele nu au fost antrenate explicit pentru cuvintele din vocabular. Pentru a putea prezice, calculăm scorul între un text reprezentând clasa dorită și imaginea de la intrare codată cu modelul vizual. Textul este de două feluri:

- word: reprezentăm clasa direct prin denumirea ei;
- photo of: reprezentăm clasa prin propoziția "*this is a photo of* ____" la finalul căreia adăugăm denumirea ei.

Utilizăm următoarele variante de arhitecturi: *(i)* pentru ResNet folosim RN50, RN101 (variantele cu 50, respectiv 101 straturi) și RN50×4, RN50×16 (variante care măresc arhitectura de 4, respectiv 16 ori din punctul de vedere al adâncimii, lungimii și rezoluției [41]); *(ii)* pentru ViT folosim ViT-B/32 și ViT-B/16 (care sunt antrenate pe patch-uri de 32×32 pixeli, respectiv 16×16 pixeli).

Tabelul 3: Evaluarea detecției de cuvinte cheie la nivel de AUPR pe setul de testare al bazei de date FlickR8K pentru trei arhitecturi—ResNet (RN), Visual Transformer (ViT), VGG—și diferite moduri de antrenare. Pentru metodele antrenate cu CLIP predicția se face comparând reprezentarea pe bază de imagine cu reprezentarea textuală a clasei, reprezentare care este de tip word sau photo of. Pentru rețeaua VGG-16 aceasta prezice direct scorul pentru fiecare cuvânt cheie în parte.

antrenare	arhitectură	word	photo of
CLIP	RN50	20.1	19.3
CLIP	RN101	20.2	20.3
CLIP	RN50×4	21.3	20.7
CLIP	RN50×16	**22.5**	21.9
CLIP	ViT-B/32	21.9	21.0
CLIP	ViT-B/16	22.2	**22.6**
COCO, FlickR30K	VGG-16	**22.8**	

Metrici de evaluare. Pentru fiecare cuvânt cheie din vocabular evaluăm aria de sub graficul precizie–reamintire (eng., *area under precision recall*; AUPR), metrică care sumarizează precizia și reamintirea (eng., *recall*). Cu cât această valoare este mai mare cu atât un sistem performează mai bine. Valoarea finală pe care o raportăm este media valorilor AUPR peste toate cele 67 de cuvinte din vocabular.

Rezultate experimentale. Prezentăm evaluarea în tabelul 3. Pentru arhitecturile antrenate CLIP observăm performanță mai bună pentru cele mai mari rețele, și anume RN50×16 și ViT-B/16 (ViT-B/16 lurează pe patch-uri mai mici decât ViT-B/32 deci este mai multă computație de efectuat). Între cele două moduri de a reprezenta clasa nu există o abordare dominantă: probabil codarea unei clase de forma photo of are sens doar pentru anumite cuvinte "*this is a photo of a ball*", dar nu se potrivește pentru verbe "*climbing*" sau adjective "*pink*". Arhitectura VGG-16 obține cea mai bună performanță, dar aceasta a fost antrenată pe baze de date similare și în condiții similare, deci acest lucru nu este surprinzător. De fapt, performanța metodelor CLIP este remarcabilă deoarece acestea operează într-un scenariu total nou fără a fi adaptate în vreun fel. Un alt aspect important, este dificultatea sarcinii care apare ca urmare a faptului că cuvintele cheie sunt extrase din descrieri textuale, ceea ce implică că acestea pot fi zgomotoase: cuvinte care apar în imagine nu e neapărat să fie descrise [3], și invers, cuvinte care nu apar în imagine e posibil să fie menționate (de exemplu, *camera* in "*a girl stands in front of the camera*").

2.2.2 Sarcina 2: Traducerea video–vorbire

În această subsecțiune discutăm cum utilizăm arhitectura vizuală propusă pentru sarcina de traducere a videourilor în audio. Mai exact, ne dorim să învățăm o mapare dintr-un videoclip silențios al unei persoane care vorbește cu semnalul de vorbire audio corespunzător. O posibilă aplicație a acestei sarcini este de a permite persoanelor care și-au pierdut capacitatea de a vorbi să interacționeze cu un sintetizator de vorbire într-un mod mai personalizat, mai rapid și mai natural. Mai multe detalii despre această metodă se găsesc în articolul nostru publicat în cadrul proiectului [22].

Setul de date. Am desfășurat experimentele pe baza de date GRID [5]—cel mai comun set utilizat pentru sarcina video–vorbire [1, 8, 43]. Setul de date constă din 34,000 de înregistrări video și audio provenite de la 34 de vorbitori diferiți. Vocabularul este limitat la 52 de cuvinte, dar nu există două înregistrări care să conțină exact aceeași secvență de cuvinte.

Detalii de implementare. Modelul de bază combină arhitectura ResNet [11] (pentru procesarea vizuală) cu decodorul Tacotron2 [32] (pentru sinteza vorbirii). Pe lângă aceste două componente, am interpus un strat de convoluții 3D și un strat de celule LSTM peste trăsăturile extrase de ResNet pentru a face modelul mai potrivit pentru procesarea unui flux video. Componenta vizuală este inițializată cu ponderile învățate pe baza de date ImageNet și primește la intrare doar zona din jurul buzelor. Rețeaua generează spectrograme Mel care sunt inversate în audio cu o rețea de super-rezoluție [40].

Metrici de evaluare. Folosim mai multe metode de evaluare pentru vorbirea sintetizate pentru a evalua diferitele fațete ale sistemelor video–vorbire, precum *calitatea* și *inteligibilitate*. În ceea ce privește *calitatea*, urmăm pe Vougioukas et al. [43] și folosim trei metrici pentru a măsura calitatea semnalului generat: distorsiunea mel-cepstrală (MCD) [15], inteligibilitatea obiectivă pe termen scurt (STOI) [39] și evaluarea perceptivă a calității vorbirii (PESQ) [29]. Pentru

Tabelul 4: Evaluarea metodelor de traducere video-la-vorbire pe baza de date GRID. Rezultatele marcate cu † indică faptul că am recalculat metricile de evaluare pe datele oferite de autori. Săgețile indică direcția în care performanța este mai bună pentru fiecare metrică.

	STOI ↑	PESQ ↑	MCD ↓	WER ↓
Lip2AudSpec [1]	0.446	1.82	38.14	32.5
V2S GAN [43]	0.518	1.71	22.29	26.6
V2S GAN [43]†	**0.525**	1.72	**22.02**	27.1
metoda de bază	0.470	**1.88**	32.28	21.8
+ informație de vorbitor	0.468	1.85	32.08	**19.9**

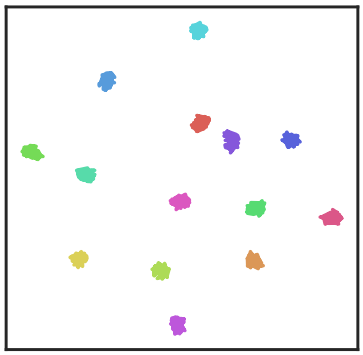

Figura 1: Proiecții t-SNE [20] ale trăsăturilor de vorbitor extrase din audio-urile sintetizate pe baza video-urilor venind de la 14 vorbitori distincți. Observăm că modelul de bază, deși agnostic la informația de vorbitor, este capabil să genereze voci distincte pentru fiecare vorbitor doar pe baza informațiilor vizuale colectate din buze.

a evalua *inteligibilitatea* ieșirii audio, utilizăm un sistem de recunoaștere automată a vorbirii pe sunetul sintetizat și raportăm rata de eroare a cuvintelor (WER).

Rezultate experimentale. Evaluăm metodele propuse pentru sarcina de traducere video–vorbire și o contrastăm cu metodele propuse anterior [1, 43]. Considerăm configurația în care se ține cont de vorbitor [43], care constă din patru vorbitori, fiecare cu 900 de mostre pentru antrenament, 50 pentru validare și 50 pentru testare. Evaluăm două variante ale metodelor noastre: modelul de bază care este independent de vorbitor și antrenat pe toate date care vin de la cei patru vorbitori simultan și un model antrenat care încorporează în mod explicit identitatea vorbitorului utilizând un vector de tip *one-hot encoding*.

Rezultatele cantitative sunt prezentate în tabelul 4, în timp ce mostre calitative pot fi găsite online: `http://speed.pub.ro/xts/`. Metodele noastre obțin rezultate competitive în ceea ce privește metricile calității vorbirii: raportăm cele mai bune rezultate pentru PESQ și cele mai bune rezultate pentru STOI și MCD. În ceea ce privește inteligibilitatea conținutului vorbirii (WER), metodele noastre dau cele mai bune rezultate. Interesant, chiar dacă modelul de bază este complet ignorant de identitatea explicită a vorbitorului, totuși încă poate produce rezultate similare cu metoda care incorporează explicit identitatea vorbitorului. Acest rezultat, împreună cu imaginea din figura 1, sugerează că rețeaua vizuală este suficient de puternică încât poate modela vorbitorii deși primește la intrare doar buzele persoanei care vorbește.

2.3 Diseminarea rezultatelor

În urma activităților de cercetare fundamentală efectuate în cadrul proiectului, au publicate două articole de revistă și unul de conferință. Astfel, au fost îndeplinite cerințele minimale de diseminare a rezultatelor pentru etapa 2 / 2021, care prevedeau publicarea unui articol de conferință. Listăm în continuare articolele finanțate prin proiectul de cercetare:

- D. Oneață and H. Cucu. Multimodal speech recognition for unmanned aerial vehicles. *Computers & Electrical Engineering*, 90:106943, 2021
- D. Oneață, A. Stan, and H. Cucu. Speaker disentanglement in video-to-speech conversion. In *29th European Signal Processing Conference*, 2021

- 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

Director proiect,
Dan Oneață

Bibliografie

[1] H. Akbari, H. Arora, L. Cao, and N. Mesgarani. Lip2AudSpec: Speech reconstruction from silent lip movements video. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2516–2520, 2018.

[2] I. Bello, W. Fedus, X. Du, E. D. Cubuk, A. Srinivas, T.-Y. Lin, J. Shlens, and B. Zoph. Revisiting ResNets: Improved training and scaling strategies. *arXiv preprint arXiv:2103.07579*, 2021.

[3] A. C. Berg, T. L. Berg, H. Daume, J. Dodge, A. Goyal, X. Han, A. Mensch, M. Mitchell, A. Sood, K. Stratos, et al. Understanding and predicting importance in images. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3562–3569, 2012.

[4] 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.

[5] M. Cooke, J. Barker, S. Cunningham, and X. Shao. An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, 120(5):2421–2424, 2006.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[7] 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.

[8] A. Ephrat and S. Peleg. Vid2speech: Speech reconstruction from silent video. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5095–5099, 2017.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] H. Kamper, G. Shakhnarovich, and K. Livescu. Semantic speech retrieval with a visually grounded model of untranscribed speech. *Transactions on Audio, Speech and Language Processing*, 27(1):89–98, 2019.

[14] 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.

[15] R. Kubichek. Mel-cepstral distance measure for objective speech quality assessment. In *IEEE Pacific Rim Conference on Communications Computers and Signal Processing*, volume 1, pages 125–128, 1993.

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

[17] 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.

[18] 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.

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

[20] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov): 2579–2605, 2008.

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

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

[23] 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.

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

[25] 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.

[26] 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.

[27] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[28] 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.

[29] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra. Perceptual evaluation of speech quality (pesq)—a new method for speech quality assessment of telephone networks and codecs. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 2, pages 749–752, 2001.

[30] 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.

[31] 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.

[32] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, et al. Natural TTS synthesis by conditioning WaveNet on Mel spectrogram predictions. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4779–4783, 2018.

[33] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[34] 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.

[35] 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.

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

[37] T. Srinivasan, R. Sanabria, F. Metze, and D. Elliott. Multimodal speech recognition with unstructured audio masking. In *Workshop on Natural Language Processing Beyond Text, EMNLP*, 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] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4214–4217, 2010.

[40] H. Tachibana, K. Uenoyama, and S. Aihara. Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention. *CoRR*, abs/1710.08969, 2017.

[41] M. Tan and Q. Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114, 2019.

[42] 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.

[43] K. Vougioukas, P. Ma, S. Petridis, and M. Pantic. Video-driven speech reconstruction using generative adversarial networks. In *Interspeech*, 2019.

[44] 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.

[45] 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.

[46] 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.

[47] 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.