# VORBIS: RECUNOAȘTEREA AUTOMATĂ A VORBIRII DIN SEMNALE MULTI-MODALE

RAPORT ȘTIINȚIFIC ȘI TEHNIC EXTINS – ETAPA 1 / 2020

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

**ABSTRACT**

În acest raport tratăm sarcina de recunoaștere automată a vorbirii bazată pe semnale multi-modale (și anume, înregistrarea audio și contextul vizual). Începem prin prezentarea sarcinii și a modului în care informația vizuală contribuie la îmbunătățirea performanței față unui sistem unimodal, bazat doar pe semnalul audio. Partea principală a raportului o constituie un studiu detaliat al stării artei și analiza comparativă lucrărilor de specialitate pe această temă. În final, plasăm sarcina de interes în contextul altor sarcini multi-modale și descriem o serie de direcții de cercetare ce pot fi trate în cadrul proiectului.

## 1 Introducere

Acest proiect tratează sarcina de recunoaștere automată a vorbirii (RAV): dată o uteranță rostită dorim transcrierea automată a acesteia în text. Deoarece în multe situații practice sunetul este în corespondență cu alte tipuri de date, ne propunem să folosim aceste informații suplimentare pentru a îmbunătăți performanța sistemului de RAV. În cadrul acestui proiect considerăm contextul vizual ca sursă adițională ce poate ajuta la dezambiguizarea cuvintelor transcrise. Ne referim în continuare la această sarcină sub numele de *recunoaștere automată a vorbirii din semnale multi-modale* (RAV-MM).

Canalul vizual poate îmbunătăți procesul de RAV atunci când ceea ce se vorbește este legat semantic de ceea ce apare în imagine; astfel de situații se întâlnesc în video-uri instrucționale, documentare, înregistrări sportive, controlul roboților de la distanță. Ca exemplu concret, putem considera cazul unui pilot care controlează o dronă de la distanță spunând "*record the skidding acitvity*" (ro., "*înregsitrează activitatea de defrișare*"); pe baza imaginii capturate de dronă (a se vedea figura 1, stânga), sistemul de RAV poate exclude ipoteze similare din punct de vedere fonetic, cum ar fi "*record the skiing activity*" (ro., "*înregistrează activitatea de schiat*"). Figura 1, dreapta, prezintă un alt exemplu care poate apărea în cazul transcrierilor video-urilor descriptive (cum ar fi video-uri instrucționale, documentare sau filme).

Recunoașterea automată a vorbirii este una dintre cele mai comune interfețe om-mașină – prin aplicațiile sale se numără asistenți inteligenți pe bază de voce (precum Amazon Alexa sau Google Home) sau diferite utilitare soft pentru persoane cu dizabilități (precum interacțiunea vocală pentru nevăzători sau generarea de subtitrări video pentru persoanele surde). Chiar dacă subiectul recunoașterii vorbirii are o lungă istorie în spate și a fost studiat intens [Jurafsky and Martin, 2008, Renals and Hain, 2010], sarcina este încă departe de a fi rezolvată și chiar și sistemele din producție sunt predispuse la erori cauzate de condiții dificile (de exemplu, zgomot, accente, reverberații) [Hanun, 2017, Szymański et al., 2020]. Din acest motiv, scopul proiectului de față este de a utiliza informații auxiliare pentru a corecta erori cauzate de astfel de scenarii.

## 2 Utilitatea canalului vizual

În introducere am prezentat exemple în care canalul vizual poate ajuta la îmbunătățirea transcrierii vorbirii. În această secțiune identificăm trei categorii în care informațiile vizuale pot ajuta la recunoașterea vorbirii. Deși această terminologie și clasificare nu este standard, am găsit-o utilă în prezentarea și înțelegerea metodelor de recunoaștere

| | | |
|---|---|---|
| audio | *record the skiing activity* | *a man shows brown linebacker knee to another* |
| audio + imagine | *record the skidding activity* | *a man jumps from a balcony to another* |

Figura 1: Exemple care ilustrează posibile beneficii aduse de utilizarea semnalului vizual într-un sistem de recunoaștere automată a vorbirii. Linia marcată cu audio indică transcrierile uteranțelor folosind doar semnalul audio; acestea conțin greșeli indicate cu roșu. Linia marcată cu audio + imagine indică rezultatul corect, obținut prin integrarea contextului vizual.

automată a vorbirii din semnale multi-modale (RAV-MM). Concret, avem următoarele trei situații în care canalul vizual este util.

**Relație simbolică** – ceea ce se vorbește este în corespondență simbolică (aproximativ de unu-la-unu) cu contextul vizual. Exemplul concret cel mai comun este cel al metodelor de citire de pe buze (eng., *lip reading*): persoana care vorbește este prezentă și vizibilă în videoclip, și există o mapare directă a mișcării buzelor la cuvintele rostite. Aceasta direcție este foarte bogată din punct de vedere al literaturii, conținând atât lucrări care încearcă transcrierea înregistrării video fără sunet în text, cât și lucrări care augmentează semnalul audio cu înregistrarea video, arătând îmbunătățirii ale transcrierii față de metode bazată doar pe semnalul audio, mai ales în condiții de zgomot. În acest raport, nu vom detalia starea artei pentru această categorie de modele multi-modale, pentru că propunerea de proiect vizează următoarele două cazuri. Remarcăm că această categorie, a relației simbolice, presupune apariția buzelor pe tot parcursul înregistrării și include ca tipuri de date discursuri sau interviuri.

**Relație semantică** – ceea ce se vorbește este corelat din punct de vedere semantic (al înțelesului) cu ceea ce este vizibil în videoclip. Exemple prezentate anterior, în introducere, și redate în figura 1 sunt cazuri ale acestui tip relație semantică între uteranță și imagine. La nivelul cel mai explicit, cuvintele rostite apar în scenă, dar această condiție nu este necesară – poate fi suficient ca acestea să se coreleze la nivel semantic cu contextul vizual, iar informațiile vizuale să ajute la adaptarea domeniului modelului de limbă. Tipul de date care exemplifică acest tip de relație includ controlul roboților de la distanță, documentare, videoclipuri instructive, comentarii sportive.

**Relație la nivel acustic** – ceea ce se vorbește este independent de canalul vizual, dar condițiile acustice se corelează cu contextul vizual. În acest caz, informațiile despre scenă pot fi utilizate pentru a adapta componenta de prelucrare acustică la condițiile de mediu. Câteva exemple concrete sunt oferite de Miao and Metze [2016], cei care au introdus această idee: dacă avem o înregistrare audio într-un tren sau în preajma unui avion, atunci vom avea pe fundal un anumit tip zgomot; pe baza acestei informații putem ajusta semnalul acustic pentru o mai bună transcriere. O altă direcție care vizează tot această relație la nivel acustic este oferită de Moriya and Jones [2019], care utilizează chipul vorbitorului pentru a extrage informații despre tipul de accent sau emoție pentru a adapta semnalul acustic. Interesant, pentru acest caz din urmă, autorii susțin că fața vorbitorului poate afecta nu doar partea acustică, cât și modelul lingvistic (de exemplu, în funcție de vârstă suntem mai predispuși de a folosi un anumit vocabular).

După cum vom vedea în secțiunea următoare, proiectarea arhitecturilor pentru sistemele de RAV-MM depind de tipul relației între cele două semnale. Acest lucru este vizibil în special pentru primele metode (cele bazate pe sistem mai clasice) în care fuziunea se face explicit fie la un nivel mai incipient (la nivelul semnalului audio, dacă vizam o relație acustică), fie mai târziu (la nivelul modelului de limbă, dacă vizam o relație semantică). Trecerea la rețele de tip end-to-end a făcut ca această clasificare să fie mai vagă pentru că a permis modele mai flexibile care fuzionează la mai multe nivele simultan. Mai mult, arhitecturile de tip end-to-end pot fi atât de flexibile încât sunt capabile să învețe oricare din cele trei tipuri de relații prezentate anterior, iar ceea ce ghidează modelul spre a învăța una din relații este tipul de date.

## 3 Metode principale

Primele abordări apar la începutul anilor 2000 și îl au ca autor pe Deb Roy [Mukherjee and Roy, 2003, Fleischman and Roy, 2008], cunoscut pentru contribuțiile pe sarcina de asociere a vorbirii și contextului vizual (eng., *grounding*) ca o metodă fundamentală de învățare pentru sistemele artificiale [Roy, 2005]. Recent, în ultimii doi-trei ani, cercetarea pe

Tabelul 1: Sumar al metodelor de recunoaștere automată a vorbirii din semnale multi-modale (audio și vizual). Pentru componenta vizuală, menționăm tipul de date utilizate în procesul de antrenare: obiecte O, scene S, acțiuni A, fețe F.

| | | vizual | | | | | | |
|---|---|---|---|---|---|---|---|---|
| referință | audio | metodă | O | S | A | F | fuziune | baze de date |
| Mukherjee and Roy [2003] | HMM-GMM | trăsături | • | | | | model limbă | descriere obiecte |
| Fleischman and Roy [2008] | HMM-GMM | trăsături | | | • | | model limbă | meciuri baseball |
| Miao and Metze [2016] | HMM-DNN | AlexNet | • | • | • | • | audio | video-uri instrucționale |
| Gupta et al. [2017] | HMM-DNN | AlexNet | • | • | | | model limbă | video-uri instrucționale |
| Sun et al. [2016] | HMM-DNN | VGG16 | • | | | | model limbă | Flickr 8K |
| Moriya and Jones [2018] | HMM-DNN | VGG19 | • | | | | model limbă | How2, Udacity |
| Palaskar et al. [2018] | end-to-end | AlexNet | • | | | | audio | How2 |
| Caglayan et al. [2019] | end-to-end | ResNet | • | • | • | | audio, codor, decodor | How2 |
| Moriya and Jones [2019] | HMM-DNN | CNN | | | | • | audio, limbă | How2, Udacity |
| Srinivasan et al. [2019] | end-to-end | ResNet | • | | | | codor, decodor | Places |
| Srinivasan et al. [2020a] | end-to-end | ResNet | • | • | | | decodor | Flickr 8K |
| Srinivasan et al. [2020b] | end-to-end | RCNN | • | | | | decodor | Flickr 8K, COCO |
| Srinivasan et al. [2020c] | end-to-end | ResNet | • | | | | codor, decodor | Flickr 8K, COCO |
| Paraskevopoulos et al. [2020] | end-to-end | ResNet | | | • | | decodor | How2 |

acest domeniu s-a intensificat, odată cu apariția de baze de date multi-modale și popularizarea rețelelor neurale adânci. Cele mai mult lucrări recente îl au în prim plan pe Florian Metze și grupul său [Miao and Metze, 2016, Gupta et al., 2017, Palaskar et al., 2018, Caglayan et al., 2019, Srinivasan et al., 2019, 2020a,b,c], care au explorat mai multe tehnici pentru această problemă. Observăm tendința domeniului de a gravita spre arhitecturi de tip *end-to-end* (după cum am propus și în acest proiect).

Un sistem de RAV-MM procesează vorbirea și contextul vizual prin două componente separate, care ulterior fuzionează pentru a combina cele două tipuri de informații. Discutăm aceste trei aspecte în continuare și încheiem cu o secțiune despre tipul de date utilizat. O prezentare comparativă a metodelor studiate este redată în Tabelul 1.

**Componenta de procesare a vorbirii.** Sistemele de recunoaștere a vorbirii au cunoscut de-a lungul timpului trei faze. Primul, și poate cel mai cunoscut model de RAV este cel de tip HMM-GMM, în care trăsăturile acustice sunt modelate folosind mixturi de Gaussiene (eng., *Gaussian mixture models*; GMM), iar tranzițiile temporale la nivel fonetic folosind modele Markov ascunse (eng., *hidden Markov models*; HMM) [Jurafsky and Martin, 2008]. Odată cu revirimentul rețelelor neurale, s-a trecut la modelul de tip HMM-DNN [Hinton et al., 2012], în care componenta GMM din tandemul HMM-GMM a fost înlocuită cu rețele neurale adânci (eng., *deep neural networks*; DNN). În ultimii ani, direcția este de a utiliza rețele adânci integral (de la un capăt la altul; eng., *end-to-end*) în tot sistemul de RAV [Hadian et al., 2018]. Modele end-to-end au câștigat în popularitate nu doar datorită performanței lor (care o egalează și chiar depășesc pe cea a RAV-urilor clasice), dar și pentru că sunt mai simple din punct de vedere conceptual și oferă un proces de antrenare unitar. Metodele de RAV-MM au urmat aceeași traiectorie; astfel, primele lucrări utilizează sisteme de tip HMM-GMM [Mukherjee and Roy, 2003, Fleischman and Roy, 2008], urmate de o serie de lucrări bazate sisteme baze pe HMM-DNN [Miao and Metze, 2016, Gupta et al., 2017, Moriya and Jones, 2018, 2019, Sun et al., 2016], iar în ultimii ani făcându-se trecerea la sisteme de tip end-to-end [Palaskar et al., 2018, Caglayan et al., 2019, Srinivasan et al., 2019, 2020a,b,c, Paraskevopoulos et al., 2020]. Cea mai populară abordare pentru arhitecturile de tip end-to-end este bazată pe rețele recurente pentru codor și decodor cuplate printr-un mecanism atenție [Palaskar et al., 2018, Caglayan et al., 2019, Srinivasan et al., 2019, 2020a,b,c]; alte variante includ rețelele recurente cu funcția de pierdere CTC [Palaskar et al., 2018] și cele care folosesc integral atenție, de tip Transformer [Paraskevopoulos et al., 2020].

**Componenta de procesare vizuală.** Comunitatea de vedere artificială a adoptat mai devreme tehnicile de tip end-to-end și acest lucru se observă și în metodele de MM-RAV: cu excepția lucrărilor de dinainte de 2010 (care utilizează tehnici clasice de procesare a imaginilor) [Mukherjee and Roy, 2003, Fleischman and Roy, 2008], toate celelalte articole folosesc rețele pentru a modela contextul vizual. Abordarea este a extrage un vector de trăsături (stratul softmax de probabilități a claselor sau stratul de dinainte, pre-softmax) care să sumarizeze informația vizuală și să fie ulterior pasat componentei de procesare a vorbirii. Este uzual ca rețele să fie pre-antrenate și în funcție de setul de date utilizat acestea pot recunoaște diferite categorii vizuale: obiecte (antrenate de obicei pe ImageNet) [Miao and Metze, 2016, Gupta et al., 2017, Palaskar et al., 2018, Srinivasan et al., 2019], scene (antrenate de obicei pe MIT Places) [Miao and Metze, 2016, Gupta et al., 2017, Caglayan et al., 2019], recunoaștere a acțiunilor (antrenate pe Kinetics) [Caglayan et al., 2019] sau fețe [Moriya and Jones, 2019]. Arhitectura tipică este bazată pe blocuri convoluționale și, de obicei, se preferă arhitecturi clasice precum AlexNet, VGG19 sau ResNet. De exemplu, Caglayan et al. [2019] folosesc diferite arhitecturi pentru diferite componente vizuale: ResNet-152 pentru obiecte, 3D ResNeXt-101 pentru acțiuni și ResNet-50 pentru

scene. În cazul în care intrarea vizuală este un video, fie se selectează un cadru aleatoriu din videoclip [Miao and Metze, 2016], fie se mediază vectorii de trăsături extrași peste mai multe cadre [Caglayan et al., 2019]. Metodele mai recente încearcă să asocieze cuvintele transcrise la regiuni din imagine [Srinivasan et al., 2020b], și pentru aceasta componenta vizuală operează la nivel de zone de imagini care conțin obiecte (cunoscute sub numele de propuneri de obiecte).

**Modul de fuziune.** Informațiile procesate de cele două componente (a vorbirii și cea vizuală) sunt combinate de sistemul de MM-RAV pentru a obține în final transcrierea semnalului audio. În metodele care utilizează HMM, există două locații predilecte de a face fuziunea celor două modalități: la nivelul semnalului audio sau la nivelul modelului de limbă. Alegerea se face în funcție de tipul datelor, mai exact în funcție de modul în care informația vizuală este utilă: dacă canalul vizual oferă informații despre mediul acustic, atunci se optează pentru o combinație la nivelul semnalului audio; dacă canalul vizual este relevant din punct de vedere semantic pentru ceea ce se rostește, atunci o combinație la nivelul modelului de limbă poate fi mai potrivită. Pentru primul tip de fuziune, cel mai simplu mod de a adapta acustic intrarea audio este prin concatenarea trăsăturilor audio cu cele vizuale Miao and Metze [2016], Palaskar et al. [2018], Moriya and Jones [2019]. O altă tehnică este *visual adaptive training* (VAT), în care trăsăturile video sunt utilizate pentru a transforma liniar trăsăturile audio $\mathbf{a}$, adică $\mathbf{a}' \leftarrow \alpha \odot \mathbf{a} + \beta$, unde coeficienții $\alpha$ și $\beta$ sunt funcții de trăsăturile video $\mathbf{v}$; această abordare este folosită de [Miao and Metze, 2016, Palaskar et al., 2018]. Pentru cel de-al doilea mod de fuziune (combinarea la nivelul modelului de limbă), se poate utiliza o tehnică similară VAT, adică actualizarea *embedding*-urilor cuvintelor folosind vectorul vizual [Gupta et al., 2017], dar abordarea mai des întâlnită rămâne prin pasarea vectorului vizual ca prim "cuvânt" în modelul de limbă [Sun et al., 2016, Moriya and Jones, 2018, Caglayan et al., 2019]; Caglayan et al. [2019] explorează și combinarea prin inițializarea stării ascunse a rețelelor recurente cu vectorul vizual. Trebuie menționat că pentru metodele bazate pe HMM fuziunea la nivelul modelului de limbă se petrece în partea de re-evaluare a ipotezelor generate de modelul de RAV (iar nu în partea de generare a ipotezelor inițiale). Bănuim că această abordare este preferată din considerente practice – este dificil de implementat decodarea folosind un model de limbă adaptat vizual. Cu toate acestea experimentele din literatură indică îmbunătățiri și în aceste condiții și, mai mult, acestea sunt departe de a fi saturate judecând după performanța unui model oracol (care alege cea mai bună ipoteză din setul inițial generat de RAV) Sun et al. [2016], Moriya and Jones [2018].

Trecerea la arhitecturile de tip end-to-end a făcut să nu mai existe o delimitare clară a modului în care canalul vizual impactează procesul de RAV. Se poate face totuși o distincție în funcție de locul unde are loc combinarea în arhitectură: dacă fuziunea este timpurie (în codificator), adaptarea afectează procesarea acustică [Caglayan et al., 2019, Srinivasan et al., 2019, 2020c], iar dacă fuziunea este târzie (în decodor), adaptarea afectează preponderent generarea transcrierii [Srinivasan et al., 2020a,b, Paraskevopoulos et al., 2020]. Deoarece metodele de tip end-to-end sunt inerent compoziționale, acestea permit mai multă variabilitate din punctul de vedere al combinațiilor. Abordări interesante în această direcție încearcă să modeleze interacțiunile dintre cele două canale (audio și imagini), majoritatea folosind atenție, de exemplu, atenție inter-modală [Srinivasan et al., 2020b, Paraskevopoulos et al., 2020], sau atenție ierarhică [Srinivasan et al., 2020c]. Srinivasan et al. [2019, 2020a,b,c] tratează situația datelor audio perturbate cu zone de zgomot în vederea augmentării procesului de antrenare și a analizei comportamentului sistemelor de RAV-MM, arătând că acestea sunt capabile să recupereze din imagini cuvintele lipsă din audio.

## 4 Seturi de date

Datele utilizate pentru această sarcină conțin trei tipuri de modalități aliniate: audio (uteranțe), vizual (imagini sau video), limbă (transcrierea). Bazele de date publice cele mai utilizate de lucrările anterioare sunt How2 [Sanabria et al., 2018] și Flickr 8K Audio [Harwath and Glass, 2015]. How2 [Sanabria et al., 2018] este un set de date cu video-uri instrucționale; aceasta are două variante, una de 300 de ore și alta de 1000 ore, dar prima varianta este cea preferată de abordările precedente, conținând 13K video-uri și 185K propoziții. Recent, a fost dezvoltată o bază de date similară, dar de scară mult mai largă, conținând peste 100M de video-uri [Miech et al., 2019]; rămâne de văzut cât este de utilă pentru proiectul de față, având în vedere că transcrierile sunt obținute automat. Flickr 8K Audio [Harwath and Glass, 2015] este un set de date derivat din Flickr 8K (o bază de date de descriere a imaginilor, conținând imagini și text) prin citirea și înregistrarea descrierilor; aceasta conține 40K uterante pentru 8K imagini. Alte seturi de date care conțin date descriptive sunt PlacesAudio [Harwath et al., 2016], cu 400K uterante rostite și derivat din setul de date Places205, și SpeechCOCO [Havard et al., 2017], cu 600K uterante sintetizate pentru 123K imagini și derivat din setul de date COCO.

## 5 Relația cu alte sarcini

Sistemele care utilizează mai multe tipuri de date senzoriale (de exemplu, audio, vizuale, text) sunt cunoscute ca sisteme multi-modale. Metodele pe care le-am prezentat anterior, pentru recunoașterea automată a vorbirii din semnale multi-modale, preiau la intrare semnalele audio și video și produc la ieșire text. Dacă considerăm însă alte combinații

| | vizual | text | audio | asocieri | sarcină (eng.) |
|---|---|---|---|---|---|
| 1 | in | out | – | – | image captioning; paragraph generation |
| 2 | out | in | – | – | text-to-image generation |
| 3 | in | out | – | out | dense image captioning, dense relational captioning |
| 4 | in | out | – | in | controllable and grounded captioning |
| 5 | in | in | – | out | phrase grounding |
| 6 | in | in + out | – | – | visual question answering; neural machine translation |
| 7 | in | in + out | – | out | referring expression recognition |
| 8 | in | – | in | out | discover visual objects and spoken words from raw sensory input |
| 9 | – | in | out | – | speech recognition |
| 10 | – | out | in | – | speech synthesis |
| 11 | out | in | – | in | image generation/retrieval from traces |
| 12 | in | out | in | in | grounded speech recognition |
| 13 | in | – | in | out | voice-driven environment navigation |

Tabelul 2: Taxonomie a sarcinilor bazate pe semnale multi-modale (vizual, text, audio) în funcție de utilizare acestora—la intrarea (*in*) sau ieșirea (*out*) sistemului. O axă suplimentară este prezența asocierilor între diferitele modalități (eng., *grounding*). Tabel adaptat din [Pont-Tuset et al., 2020].

de intrare-ieșire pentru aceste tipuri de semnale, obținem noi sarcini de învățare care prin metodologia aplicată pot fi relevante și pentru problema noastră; tabelul 2 enumeră sarcini de acest tip.

### 5.1 Metode audio-vizuale

O direcție foarte populară care implică semnale multi-modale este cea care utilizează semnale audio și vizuale fără a face uz de adnotările la nivel de text (linia 8 în tabelul 2). Motivația acestei sarcini este că acest tip de date este mult mai ușor de procurat decât cel care implică și text, pentru că acesta din urmă necesită adesea adnotări suplimentare. Ideea acestor metode este de a învăța corelații între cele două tipuri de intrări pentru descoperirea de noi concepte sau pentru sarcina de regăsire a informațiilor (eng., *information retreival*).

Concret, sistemele audio-vizuale vizează sarcini precum regăsirea imaginilor pe baza vorbirii [Synnaeve et al., 2014, Harwath et al., 2016, 2018], învățarea de reprezentări numerice distribuite (eng., *embeddings*) [Harwath et al., 2016, Harwath and Glass, 2015], localizarea obiectelor pe baza vorbirii [Harwath et al., 2018] sau detecția semantică a cuvintelor cheie [Kamper et al., 2019]. Abordarea tipică exploatează corespondențele statistice și învață reprezentări numerice pentru cele două modalități, text și imagini, într-un sub-spațiu comun. Lucrările lui Harwath și Glass sunt printre primii exponenți ai acestei direcții; prima lor abordare s-a bazat pe corespondențe extrase la un nivel grosier (segmente de vorbire și detecții de obiecte) [Harwath and Glass, 2015], în timp ce tehnicile ulterioare învață corelații la rezoluții mai fine (cadre de vorbire și pixeli de imagine) [Harwath et al., 2018]. Diferit, Kamper et al. [2019] valorifică informațiile vizuale pentru a antrena un sistem de identificare a cuvintelor cheie. Abordarea lor se bazează pe un ansamblu de rețele de tip elev-profesor: rețeaua profesor funcționează pe semnalul vizual și este antrenată pentru detectarea obiectelor folosind seturi de date de imagini, în timp ce rețeaua elev primește vorbire ca intrare și prezice cuvintele cheie rostite.

Abordările audio-vizuale pot fi utile pentru sarcina noastră de interes în cel puțin două moduri. Primul și cel mai direct mod este de a integra textul ca o intrare suplimentară în aceste tipuri de arhitecturi; la inferență putem recunoaște vorbirea prin regăsirea textului tratând problema ca una de regăsire a informațiilor [Rouditchenko et al., 2020]. Cel de-al doilea mod constă în învățarea de reprezentări mai bune (folosind tehnici de auto-învățare, eng., *self-supervised*) și ulterior utilizarea acestor reprezentări pentru sarcina de recunoaștere a vorbirii [Hsu et al., 2019].

### 5.2 Metode multi-modale de traducere automată

Sarcina de traducere automată folosind semnale multi-modale (TA-MM) constă în utilizarea textului și informațiilor vizuale la intrare, și producere textului (în limba țintă) la ieșire (linia 6 în tabelul 2). Această sarcina este similară cu cea de recunoaștere automată a vorbirii din semnale multi-modale (RAV-MM), doar că intrarea audio este înlocuită cu text (informație tot de tip secvențial). Diferența principală este între relația dintre intrare și ieșire: pentru RAV-MM legătura între audio și text este sintactică (vorbirea este direct corelată cu textul rostit), pentru TA-MM legătura între textul în limba sursă și textul în limba țintă este semantică (ambele texte au același înțeles, dar pot lua forme diferite). Din acest motiv, incorporarea semnalelor vizuale (care se pretează bine la legături semantice după cum am discutat în secțiunea 2) poate fi chiar mai utilă în cazul sarcinii de TA-MM.

În pofida diferențelor dintre cele două sarcini, arhitecturile folosite pentru TA-MM pot fi folosite și pentru sarcina noastră de interes. Menționăm în continuare trei idei care au fost utilizate cu succes pentru TA-MM:

**Doubly attention** [Calixto et al., 2017] sunt arhitecturi de tip codor-decodor care integrează cele două modalități de intrare în decodor prin utilizarea alternativă a atenției peste text, respectiv imagini. Acest arhitectură diferă de modulul de cross-atenție folosit de Paraskevopoulos et al. [2020] care fuzionează ambele modalități înainte de decodor.

**Imagination** [Elliott and Kádár, 2017] reprezintă un mod de învățare de tip multi-sarcină (eng., *multi-task learning*). Ideea este de a decupla sarcina de TA-MM în două sarcini: *(i)* traducere uni-modală din limba sursă în limba țintă; *(ii)* generare de trăsături folosind text din limba sursă.

**Deliberation networks** [Xia et al., 2017] reprezintă augmentarea unei rețele cu o etapă suplimentară de decodare; astfel, o rețea de tip secvență-la-secvență va avea două stagii de decodare: unul preliminar și cel de-al doilea de rafinare, care folosește informația globală de la prima etapă pentru a îmbunătății performanța. Intuitiv, această metodă adaugă un pas de re-evaluare (eng., *rescoring*; uzual în metodele de recunoaștere automată a vorbirii), dar într-un mod integrat (de tip end-to-end).

O serie de lucrări analizează utilitatea canalului vizual în sarcina de traducere automată. De exemplu, Frank et al. [2018] au recurs la translatori profesioniști pentru a evalua rolul contextului vizual în procesul de traducere; concret, traducătorii au corectat din traducerile bazate doar pe text folosind suplimentar imaginea aferentă textului. Rezultatele indică faptul că doar în 2.2% (pe partea de `dev`), respectiv 1% (pe partea de `test`) a fost necesar ca adnotatorii să intervină; în 30% din aceste situații, problema identificată fiind descrierea greșită a imaginii. Rezultatele nu par încurajatoare la prima vedere, dar trebuie menționat că experimentul a pornit de la traduceri de o calitate înaltă (obținute de la translatori), în timp ce în cazul traducerilor automate performanța de la care pornim este mai joasă și canalul vizual poate ajuta mai mult. Recent, Wu et al. [2021] analizează ponderile oferite canalului vizual și observă că aceste valori sunt mici, concluzionând astfel că modelele multi-modale nu fac uz de canalul vizual, deși acestea aduc îmbunătățiri față de modelul unimodal. În continuare, autorii observă că aceste îmbunătățiri sunt similare cu cele obținute prin regularizarea modelelor și consideră acesta ca un argument pentru faptul că ramura vizuală din arhitectură joacă rol de regularizare. O observație similară a fost făcută și de Srinivasan et al. [2019] în contextul sistemelor de RAV-MM.

## 6 Direcții de lucru

Din momentul depunerii proiectului (octombrie 2019), direcția a avansat notabil, noi identificând cel puțin șapte articole foarte relevante în ultimii doi ani (vezi tabelul 1). În lumina acestor ultime lucrări din domeniu, considerăm următoarele piste de lucru.

- *Analiza îmbunătățirilor obținute prin canalul vizual.* Un pas important în înțelegerea metodelor de RAV-MM este evaluarea utilității informației vizuale. O aproximare a utilității poate fi obținută prin corectarea automată a greșelilor de transcriere ale unui sistem de RAV unimodal folosind un model multi-modal de limbă mascat (eng., *masked language model*) [Lu et al., 2019, Tan and Bansal, 2019] – acesta ar primi la intrare transcrierea (cu cuvintele greșite mascate) și contextul vizual și ar scoate la ieșire cuvintele corectate. Dacă un astfel de procedeu reușește să corecteze multe erori, putem concluziona că informația multi-modală este benefică. Această analiză reprezintă varianta automată a celei prezentate în [Frank et al., 2018] și ar putea fi utilizată pentru mai multe tipuri de date pentru a investiga care din ele se pretează mai bine la această sarcina de RAV-MM.
- *Factorizarea în sub-probleme.* Sistemele de RAV-MM necesită pentru antrenare triple de tip audio, imagine, text. Din punct de vedere practic, acest tip de asocieri poate fi dificil de procurat, o situație mai fezabilă și posibil mai realistă reprezentându-o utilizarea de asocieri factorizate: de exemplu, *(i)* perechi audio–imagini și perechi audio–text sau *(ii)* perechi audio–imagini și text separat. De asemenea, cantitățile de date pentru fiecare din aceste asocieri diferă; de exemplu, textul este cu siguranță mai comun decât perechi audio–imagini. Setări experimentale similare au fost explorate și de Elliott and Kádár [2017] sau Zhang et al. [2019], dar pentru sarcina de traducere multi-modală.
- *Utilizarea adnotărilor structurate.* Tendința recentă din domeniul învățării automate este de a recurge la mai puțină supervizie, cu scopul de a putea utiliza cât mai multe date. Cu toate acestea, în cazul în care avem la dispoziție date deja adnotate este interesant de explorat direcția opusă, și anume, cum putem utiliza adnotări cât mai bogate pentru a rezolva sarcini mai deosebite. Seturi de date care oferă adnotări mai structurate sunt SpeechCOCO [Havard et al., 2017] sau Localized Narratives [Pont-Tuset et al., 2020]; aceste baze date oferă imagini, audio, descrieri textuale și localizarea acestora în timp și spațiu, diferența fiind despre cum

este reprezentată asocierea text–imagine: în SpeechCOCO aceasta este făcută prin segmentări ale obiectelor relevante, în Localzed Narratives prin trasarea cu unor regiuni cu mouse-ul. Aceste date ne oferă posibilitatea de a augmenta un sistem de RAV-MM cu localizări ale transcrierilor.

- *Integrarea de cunoștințe anterioare în arhitecturi.* Majoritatea arhitecturilor curente pentru RAV-MM sunt generice. În principiu, flexibilitatea lor este de folos când avem cantități foarte mari de date, dar dacă dorim ca acestea să fie mai eficiente din punctul de vedere al datelor trebuie să incorporăm cunoștințe anterioare (eng., *prior knowledge* sau eng., *inductive bias* [Battaglia et al., 2018]) în arhitecturile utilizate. Un exemplu concret este de a utiliza un mecanism de gating vizual la nivel de cuvânt pentru a modela faptul că doar anumite cuvinte au context vizual . Acesta reprezintă un pas în direcțiile explorate de Wu et al. [2021] (care folosesc un mecanism de gating vizual, dar la nivel global, de propoziție) și Paraskevopoulos et al. [2020] (care folosesc un mecanism de ponderare, dar fix pentru toate exemplele).

## 7 Concluzii

În acest raport am prezentat și comparat cele mai relevante articole pentru sarcina de recunoaștere automată a vorbirii din semnale mult-modale. Literatura de specialitate indică faptul că aceasta este o direcție în continuă explorare, numai în ultimii doi ani fiind publicate șapte articole foarte relevante pe această temă. Tendința este de a utiliza rețele de tip end-to-end care să integreze ambele semnale (audio și imagini) la mai multe nivele și de a încerca 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.

## Bibliografie

P. W. Battaglia, J. B. Hamrick, V. Bapst, A. Sanchez-Gonzalez, V. Zambaldi, M. Malinowski, A. Tacchetti, D. Raposo, A. Santoro, R. Faulkner, C. Gulcehre, F. Song, A. Ballard, J. Gilmer, G. Dahl, A. Vaswani, K. Allen, C. Nash, V. Langston, C. Dyer, N. Heess, D. Wierstra, P. Kohli, M. Botvinick, O. Vinyals, Y. Li, and R. Pascanu. Relational inductive biases, deep learning, and graph networks. *ArXiv e-prints*, 6 2018.

Ozan Caglayan, Ramon Sanabria, Shruti Palaskar, Loic Barraul, and Florian Metze. Multimodal grounding for sequence-to-sequence speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8648–8652, 2019.

Iacer Calixto, Qun Liu, and Nick Campbell. Doubly-attentive decoder for multi-modal neural machine translation. In *Association for Computational Linguistics*, pages 1913–1924, 2017.

Desmond Elliott and Ákos Kádár. Imagination improves multimodal translation. In *International Joint Conference on Natural Language Processing*, pages 130–141, 2017.

Michael Fleischman and Deb Roy. Grounded language modeling for automatic speech recognition of sports video. In *Association for Computational Linguistics*, pages 121–129, 2008.

Stella Frank, Desmond Elliott, and Lucia Specia. Assessing multilingual multimodal image description: Studies of native speaker preferences and translator choices. *Natural Language Engineering*, 24(3):393–413, 2018.

Abhinav Gupta, Yajie Miao, Leonardo Neves, and Florian Metze. Visual features for context-aware speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5020–5024, 2017.

Hossein Hadian, Hossein Sameti, Daniel Povey, and Sanjeev Khudanpur. End-to-end speech recognition using lattice-free MMI. In *Interspeech*, pages 12–16, 2018.

Awni Hanun. Speech recognition is not solved. https://awni.github.io/speech-recognition/, 2017. Accessed: 2020-12-09.

David Harwath and James Glass. Deep multimodal semantic embeddings for speech and images. In *Workshop on Automatic Speech Recognition and Understanding*, pages 237–244, 2015.

David Harwath, Antonio Torralba, and James Glass. Unsupervised learning of spoken language with visual context. In *Advances in Neural Information Processing Systems*, pages 1858–1866, 2016.

David Harwath, Adrià Recasens, Dídac Surís, Galen Chuang, Antonio Torralba, and James Glass. Jointly discovering visual objects and spoken words from raw sensory input. In *European Conference on Computer Vision*, pages 649–665, 2018.

William Havard, Laurent Besacier, and Olivier Rosec. SPEECH-COCO: 600k visually grounded spoken captions aligned to MSCOCO data set. In *Workshop on Grounding Language Understanding*, pages 42–46, 2017.

Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-Rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, and Tara N Sainath. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97, 2012.

Wei-Ning Hsu, David Harwath, and James Glass. Transfer learning from audio-visual grounding to speech recognition. In *Interspeech*, pages 3242–3246, 2019.

Daniel Jurafsky and James H Martin. *Speech and language processing*. Prentice Hall, 2008.

Herman Kamper, Gregory Shakhnarovich, and Karen Livescu. Semantic speech retrieval with a visually grounded model of untranscribed speech. *Transactions on Audio, Speech and Language Processing*, 27(1):89–98, 2019.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pages 13–23, 2019.

Yajie Miao and Florian Metze. Open-domain audio-visual speech recognition: A deep learning approach. In *Interspeech*, pages 3414–3418, 2016.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. HowTo100M: Learning a text-video embedding by watching hundred million narrated video clips. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2630–2640, 2019.

Yasufumi Moriya and Gareth JF Jones. LSTM language model adaptation with images and titles for multimedia automatic speech recognition. In *IEEE Spoken Language Technology Workshop*, pages 219–226, 2018.

Yasufumi Moriya and Gareth JF 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.

Niloy Mukherjee and Deb Roy. A visual context-aware multimodal system for spoken language processing. In *European Conference on Speech Communication and Technology*, 2003.

Shruti Palaskar, Ramon Sanabria, and Florian Metze. End-to-end multimodal speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5774–5778, 2018.

Georgios Paraskevopoulos, Srinivas Parthasarathy, Aparna Khare, and Shiva Sundaram. Multiresolution and multimodal speech recognition with transformers. In *Association for Computational Linguistics*, 2020.

Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives. In *European Conference on Computer Vision*, pages 647–664, 2020.

Steve Renals and Thomas Hain. Speech recognition. In A. Clark, C. Fox, and S. Lappin, editors, *Computational Linguistics and Natural Language Processing Handbook*. Blackwells, 2010.

Andrew Rouditchenko, Angie Boggust, David Harwath, Dhiraj Joshi, Samuel Thomas, Kartik Audhkhasi, Rogerio Feris, Brian Kingsbury, Michael Picheny, Antonio Torralba, et al. AVLnet: Learning audio-visual language representations from instructional videos. *arXiv preprint arXiv:2006.09199*, 2020.

Deb Roy. Grounding words in perception and action: Computational insights. *Trends in cognitive sciences*, 9(8): 389–396, 2005.

Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. How2: A large-scale dataset for multimodal language understanding. In *Advances in Neural Information Processing Systems*, 2018.

Tejas Srinivasan, Ramon Sanabria, and Florian Metze. Analyzing utility of visual context in multimodal speech recognition under noisy conditions. In *The How2 Challenge: New Tasks for Vision & Language, ICML*, 2019.

Tejas Srinivasan, Ramon Sanabria, and Florian Metze. Looking enhances listening: Recovering missing speech using images. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6304–6308, 2020a.

Tejas Srinivasan, Ramon Sanabria, Florian Metze, and Desmond Elliott. Fine-grained grounding for multimodal speech recognition. In *Empirical Methods in Natural Language Processing*, 2020b.

Tejas Srinivasan, Ramon Sanabria, Florian Metze, and Desmond Elliott. Multimodal speech recognition with unstructured audio masking. In *Workshop on Natural Language Processing Beyond Text, EMNLP*, 2020c.

Felix Sun, David Harwath, and James Glass. Look, listen, and decode: Multimodal speech recognition with images. In *IEEE Spoken Language Technology Workshop*, pages 573–578, 2016.

Gabriel Synnaeve, Maarten Versteegh, and Emmanuel Dupoux. Learning words from images and speech. In *NIPS Workshop on Learning Semantics*, 2014.

Piotr Szymański, Piotr Żelasko, Mikolaj Morzy, Adrian Szymczak, Marzena Żyła-Hoppe, Joanna Banaszczak, Lukasz Augustyniak, Jan Mizgajski, and Yishay Carmiel. WER we are and WER we think we are. In *Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2020.

Hao Tan and Mohit Bansal. LXMERT: Learning cross-modality encoder representations from transformers. In *International Joint Conference on Natural Language Processing*, pages 5103–5114, 2019.

Zhiyong Wu, Lingpeng Kong, and Ben Kao. Good for misconceived reasons: Revisiting neural multimodal machine translation, 2021. URL https://openreview.net/forum?id=Q9U_H8lQ4yV.

Yingce Xia, Fei Tian, Lijun Wu, Jianxin Lin, Tao Qin, Nenghai Yu, and Tie-Yan Liu. Deliberation networks: Sequence generation beyond one-pass decoding. In *Advances in Neural Information Processing Systems*, pages 1784–1794, 2017.

Zhuosheng Zhang, Kehai Chen, Rui Wang, Masao Utiyama, Eiichiro Sumita, Zuchao Li, and Hai Zhao. Neural machine translation with universal visual representation. In *International Conference on Learning Representations*, 2019.