Zofia Lubańska[1], Gabriel Tarasiuk[2]

Akademia Bialska im. Jana Pawła II, Biała Podlaska

https://doi.org/10.29316/9788368103304_10

AUTOMATYCZNE ROZPOZNAWANIE MOWY Z RUCHU UST PRZY UŻYCIU SZTUCZNEJ INTELIGENCJI – PRZEGLĄD METOD, ZASTOSOWAŃ I WYZWAŃ

AUTOMATIC SPEECH RECOGNITION FROM LIP MOVEMENTS USING ARTIFICIAL INTELLIGENCE – AN OVERVIEW OF METHODS, APPLICATIONS, AND CHALLENGES

Streszczenie

Rozpoznawanie mowy na podstawie ruchów warg (znane również jako czytanie z ruchu warg) stanowi kluczowy obszar na styku wizji komputerowej i przetwarzania mowy. Technologia ta umożliwia zrozumienie mowy poprzez analizę ruchów warg i języka mówiącego bez wykorzystania dźwięku. Jest ona szczególnie ważna dla osób z ubytkiem słuchu oraz w sytuacjach, w których sygnał audio jest niedostępny lub zakłócony przez hałas. W ostatnich latach poczyniono znaczne postępy w automatycznym odczytywaniu z ruchu warg dzięki rozwojowi sztucznej inteligencji, a zwłaszcza głębokich sieci neuronowych. Niniejszy rozdział zawiera przegląd aktualnych badań nad rozpoznawaniem

[1] e-mail: z.lubanska@dyd.akademiabialska.pl

[2] student, e-mail: tarasiuk22914@stud.akademiabialska.pl

mowy na podstawie ruchu warg z wykorzystaniem sztucznej inteligencji. Przedstawiono ewolucję podejść, od tradycyjnych metod opartych na cechach wizualnych i modelach statystycznych po nowoczesne architektury głębokiego uczenia (CNN, RNN, Transformer). Omówiono typowe metody i procesy rozpoznawania, a także popularne zbiory danych (takie jak GRID, LRW i LRS) oraz porównano wydajność wybranych modeli na tych zbiorach danych.

Słowa kluczowe: wizualne rozpoznawanie mowy, czytanie z ruchu warg, sieci neuronowe, głębokie uczenie, multimodalne rozpoznawanie mowy

Summary

Speech recognition based on lip movement (also known as lip reading) is a key area at the intersection of computer vision and speech processing. It enables understanding of speech by analysing the movements of the speaker's movements of the lips and other visible articulatory features without using sound. This technology is particularly important for individuals with hearing loss and in situations where the audio signal is unavailable or distorted by noise. In recent years, significant progress has been made in automatic lip reading, thanks to advances in artificial intelligence, particularly deep neural networks. This chapter provides an overview of current research on AI-based speech recognition from lip movements. The evolution of approaches is presented, from traditional methods based on visual features and statistical models to modern deep learning architectures (CNN, RNN, Transformer). Typical recognition methods and pipelines, as well as popular datasets (such as GRID, LRW, and LRS), are discussed, and the performance of selected models on these datasets is compared.

Keywords: visual speech recognition, lip reading, neural networks, deep learning, multimodal speech recognition

Wstęp

Komunikacja werbalna to złożony proces multimodalny, w którym poza dźwiękiem mowy istotną rolę odgrywają sygnały wizualne – ruchy ust, mimika twarzy, gesty itp. (Battista i in., 2025). Ludzie intuicyjnie wykorzystują informację wizualną podczas słuchania, zwłaszcza w warunkach hałasu. Zdolność czytania z ruchu ust (ang. *lip-reading*, nazywana też wizualnym rozpoznawaniem mowy) odgrywa ważną rolę w rozumieniu mowy przy zaburzeniach słuchu lub

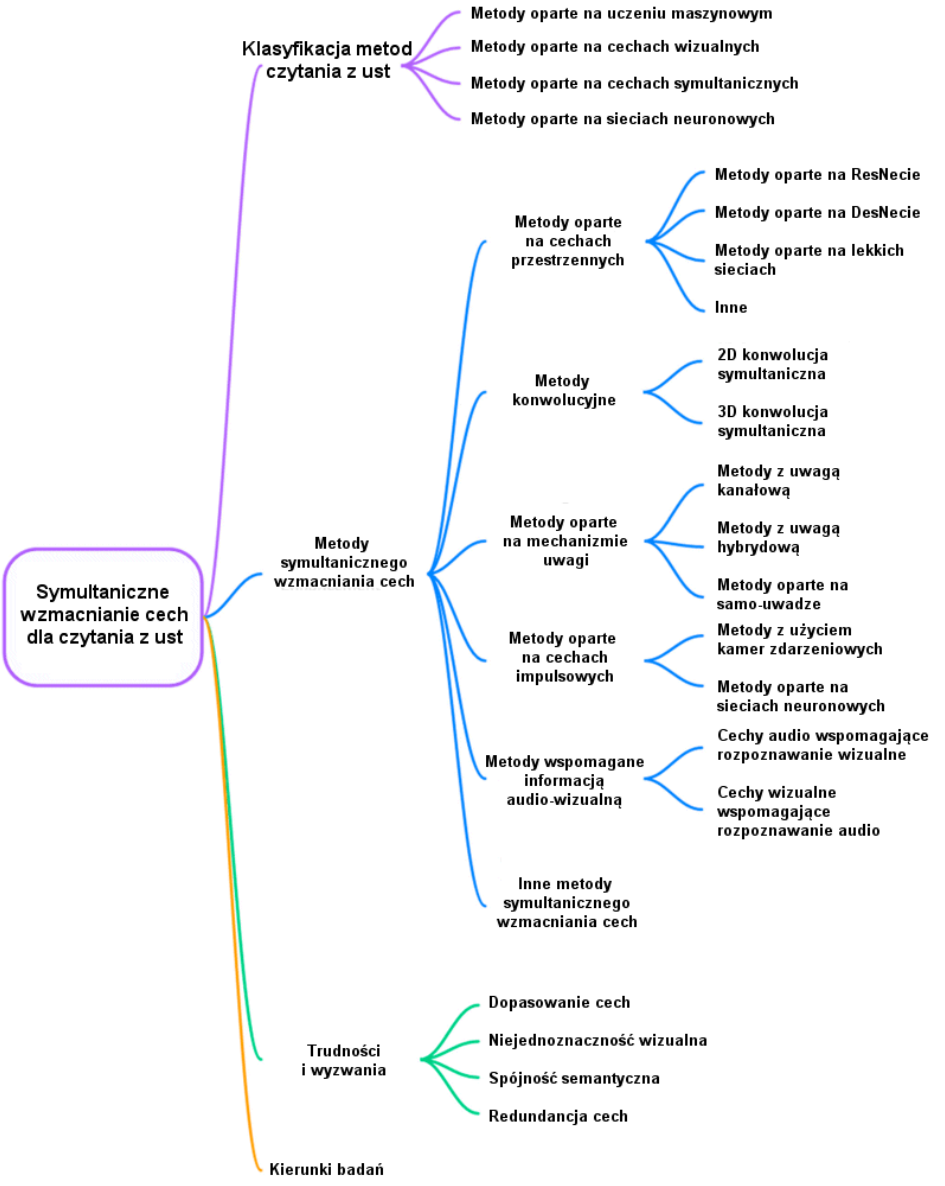
zakłóceniach akustycznych (Battista i in., 2025; Santana i in., 2025). U osób z ubytkiem słuchu umiejętność czytania z ust często staje się podstawowym kanałem komunikacji, znacząco poprawiając rozumienie przekazu (Santana i in., 2025). Również w codziennych sytuacjach, gdy np. rozmawiamy w hałaśliwym otoczeniu albo widzimy mówiącą osobę z daleka za szklaną przegrodą, analizujemy ruch ust, by wspomóc percepcję mowy (Battista i in., 2025). Nic więc dziwnego, że od dekad trwają prace nad zautomatyzowaniem tej umiejętności – tak, by komputery potrafiły "słyszeć oczami", czyli rozpoznawać mowę wyłącznie na podstawie obrazu twarzy mówiącego.

Automatyczne rozpoznawanie mowy z ruchu ust jest wyzwaniem interdyscyplinarnym, leżącym na pograniczu przetwarzania obrazów i rozpoznawania mowy (Potamianos i in., 2003; Rahmatullah i in., 2025). Tradycyjnie do tego zadania stosowano podejścia oparte na analizie cech wizualnych i klasycznych modelach uczenia maszynowego. Już na początku lat 90. badacze zademonstrowali pierwsze systemy rozpoznające proste wypowiedzi na podstawie nagrań wideo – przykładowo Mase i Pentland (1991) zastosowali analizę przepływu optycznego do śledzenia ruchu warg i rozpoznawania głosek. Pierwsze te systemy działały jednak w ograniczonym zakresie słów i wymagały ręcznej inżynierii cech. W kolejnych latach rozwijano metody ekstrakcji cech wizualnych z okolic ust, takie jak kształt konturu warg, sygnały kinematyczne czy transformacje obrazu (np. współczynniki DCT, DWT, modele kształtu) (Matthews i in., 2002; Lin i in., 2017; Miled i in., 2022). Cechy te zasilano klasyczne klasyfikatory sekwencyjne (ukryte modele Markowa HMM, modele k-NN, SVM itp.) do rozpoznawania ciągów głosek czy słów (Potamianos i in., 2003; Erbey, Barışçı, 2022). Mimo pewnych sukcesów (np. Matthews i wsp. (2002) osiągnęli ok. 63% trafności rozpoznawania cyfr na podstawie analizy kształtu ust), tradycyjne podejścia borykały się z wieloma ograniczeniami – małą odpornością na zmienność mówców, oświetlenia, kąt kamery oraz z fundamentalnym problemem niejednoznaczności wizualnej wielu fonemów (tzw. wizemy – różne dźwięki mowy mogą wyglądać tak samo na ustach) (Bear, Harvey, 2017; Rahmatullah i in., 2025). Przykładowo, głoski /p/, /b/, /m/ mają identyczny układ warg (jednostka wizualna), przez co system bazujący tylko na obrazie nie odróżni ich bez dodatkowego kontekstu. To tzw. zjawisko wizem, które utrudnia rozpoznawanie mowy wyłącznie z ruchu ust.

Dopiero rozwój głębokiego uczenia na przełomie ostatniej dekady przyniósł przełom w omawianej dziedzinie (Fenghour i in., 2021; Erbey, Barışçı, 2022; Rahmatullah i in., 2025). Pojawienie się dużych zbiorów danych wideo oraz technik głębokich sieci neuronowych (konwolucyjnych CNN i rekurencyjnych RNN/LSTM) pozwoliło budować modele automatycznie uczące się reprezentacji ruchu ust, znacznie przewyższające wcześniejsze rozwiązania (Fenghour i in., 2021; Wang i in., 2022). Na Rysunku 1 zebrano schematycznie technologie czytania z ust i ich rozwój (Po lewej wyróżniono tradycyjne podejścia *machine learning* (bazujące na ręcznie projektowanych cechach wizualnych i klasyfikatorach, np. HMM, SVM)). Po prawej pokazano podział metod głębokiego uczenia ze względu na typ przetwarzanych cech. W 2016 r. grupa z Oxford University zaprezentowała zestaw danych LRW obejmujący 500 słów wypowiadanych przez setki mówców w warunkach zbliżonych do rzeczywistych (materiały z programów BBC). Już dwa lata później najlepsze modele osiągnęły na nim ~83% skuteczności rozpoznawania słów, podczas gdy pierwsze próby dekady wcześniej dawały wyniki rzędu 50-60% (Fenghour i in., 2021). W kolejnych latach powstały jeszcze większe bazy, m.in. LRS2 i LRS3 z tysiącami pełnych zdań w języku angielskim oraz LRW-1000 – obejmujący 1000 słów w języku mandaryńskim (718 tysięcy próbek od >2000 mówców) (Wang i in., 2022; Rahmatullah i in., 2025). Dzięki temu możliwe stało się trenowanie end-to-end głębokich modeli, które automatycznie uczą się zarówno optymalnych cech wizualnych, jak i modelu języka (sekwencji). Rezultaty są imponujące – np. model LipNet (Assael i in., 2017) osiągnął 95% trafności rozpoznawania pełnych zdań w korpusie GRID, co przewyższyło wynik doświadczonych ludzkich "czytaczy z ust" (~52%). Również w zadaniach rozpoznawania pojedynczych słów w trudnych warunkach tzw. "in the wild" notuje się dziś dokładność przekraczającą 90% (Fenghour i in., 2021). Mimo to automatyczne czytanie z ust nadal pozostaje wyzwaniem. W warunkach rzeczywistych skuteczność spada – na dużym zestawie LRW-1000 (język chiński) najlepsze obecnie modele osiągają tylko ~45% poprawnych rozpoznań (Wang i in., 2022). Ponadto modele wymagają olbrzymich zbiorów danych treningowych i miewają problemy z generalizacją do nowych mówców czy języków (Rahmatullah i in., 2025).

Celem niniejszego rozdziału jest przedstawienie przeglądu stanu badań nad rozpoznawaniem mowy z ruchu ust z wykorzystaniem sztucznej inteligencji. Omówiono główne podejścia techniczne, począwszy od klasycznych metod

machine learning, aż po współczesne głębokie sieci neuronowe, wraz z charakterystyką architektur sieci oraz sposobami reprezentacji ruchu ust.



Rysunek 1. Taksonomia metod rozpoznawania mowy z ruchu ust z punktu widzenia rodzaju wykorzystywanych cech i architektur

Źródło: opracowanie własne na podstawie: Erbey, Barışçı, 2022; Fenghour i in., 2021.

W dalszej części zestawiono i porównano wybrane rozwiązania, przedstawiając dwie tabelaryczne syntezy: (a) porównanie najważniejszych modeli AI do czytania z ust (architektury i dokładności) oraz (b) porównanie właściwości najczęściej używanych zbiorów danych. Następnie omówiono praktyczne zastosowania technologii, bieżące ograniczenia i otwarte problemy oraz wskazano perspektywiczne kierunki dalszych badań.

Metody

Automatyczne rozpoznawanie mowy z ruchu ust można ogólnie podzielić na dwa etapy: (1) ekstrakcja cech wizualnych z sekwencji obrazów przedstawiających usta/twarz mówiącego, oraz (2) modelowanie sekwencji cech w celu zdekodowania treści mowy (np. klasyfikacja do słów lub transkrypcja zdania) (Potamianos i in., 2003; Fenghour i in., 2021). Historycznie najpierw dominowały podejścia szkieletowe (ang. handcrafted) – ręcznie projektowane cechy i klasyfikatory. Obecnie standardem są metody *end-to-end* z głębokimi sieciami neuronowymi, które uczą się reprezentacji i dekodowania jednocześnie. W niniejszym rozdziale przedstawiono najważniejsze techniki z obu tych paradygmatów.

Metody klasyczne (machine learning)

W początkowym okresie badań nad automatycznym czytaniem z ust (lata 90. i pocz. 2000) stosowano głównie podejścia oparte na klasycznym uczeniu maszynowym bez głębokich sieci (Potamianos i in., 2003). Kluczowym zadaniem była inżynieria dobrych cech wizualnych opisujących układ ust w kolejnych klatkach wideo. Wykorzystano szereg technik: cechy geometryczne (np. wysokość i szerokość otwarcia ust, kształt konturu warg), cechy opisujące teksturę w obszarze ust (np. współczynniki transformat – dyskretnej kosinusowej DCT lub falkowej DWT – obliczone z obrazu ust), cechy ruchu (wektory przepływu optycznego w regionie ust) oraz modele parametryczne (np. aktywne modele kształtu – AAM, śledzące kształt warg) (Matthews i in., 2002; Lin i in., 2017; Miled i in., 2022). Tabela 1 (pierwszy wiersz) zestawia typowe metody ekstrakcji cech stosowane w tych podejściach.

Następnie sekwencje takich wektorów cech z kolejnych klatek były klasyfikowane za pomocą sekwencyjnych modeli statystycznych. Najpopularniejsze były ukryte modele Markowa (HMM) – analogicznie jak w klasycznym rozpoznawaniu mowy audio – a także modele k-NN, GMM czy sieci neuronowe o płytkiej architekturze (Potamianos i in., 2003; Erbey, Barışçı, 2022). Przykładowo, znana praca Matthews i wsp. (2002) użyła modelu AAM do śledzenia ruchu ust i ekstrakcji parametrów kształtu oraz intensywności, a następnie zastosowała klasyfikator HMM do rozpoznawania głosek i słów – osiągając ~62% dokładności dla słów i ~79% dla prostych fraz na zbiorze MIRACL-VC1. Inna ówczesna metoda (Zhou i in., 2011, opisana m.in. Erbey, Barışçı, 2022) wykorzystała transformację Random Forest Manifold Alignment do dopasowania cech wizualnych różnych mówców przed klasyfikacją. Ogólnie jednak skuteczność klasycznych metod była umiarkowana, a każdy krok (wykrywanie ust, ekstrakcja cech, rozpoznawanie) wymagał osobnego dostrajania.

Dużym wyzwaniem okazała się wspomniana wcześniej niejednoznaczność fonetyczno-wizualna – wiele fonemów mapuje się na ten sam wizem (np. głoski {p, b, m} wyglądają identycznie). Rozwiązania próbowano szukać w definiowaniu odpowiednich mapowań fonem-wizem, które grupują dźwięki w klasy wizualne. Bear i Harvey (2017) dokonali obszernej analizy takich mapowań, pokazując ich wpływ na skuteczność rozpoznawania – nawet najlepsze mapowania wprowadzają jednak pewne straty informacji (Bear, Harvey, 2017). Podsumowując, tradycyjne podejścia *machine learning* stworzyły podwaliny automatycznego czytania z ust, ale miały liczne ograniczenia. Dopiero pojawienie się głębokich sieci neuronowych pozwoliło znacząco podnieść dokładność dzięki automatycznemu odkrywaniu złożonych cech czasowo-przestrzennych z danych.

Metody oparte na głębokim uczeniu

Obecnie dominującą rolę w rozpoznawaniu mowy z ruchu ust odgrywają metody wykorzystujące głębokie sieci neuronowe (Fenghour i in., 2021; Erbey, Barışçı, 2022; Deshpande i in., 2025). Cechuje je podejście *end-to-end*: surowe dane wizyjne (sekwencje obrazów) są przekazywane przez sieć, która uczy się zarówno reprezentacji cech, jak i modelowania ich zmian w czasie, a na wyjściu bezpośrednio generuje decyzję – rozpoznane słowo, zdanie albo transkrypcję

fonetyczną. W tej sekcji omawiamy główne typy architektur sieci stosowanych w literaturze, zgodnie z podziałem zaznaczonym na Rysunku 1.

Architektury CNN+RNN (cechy wizualne + cechy temporalne)

Pierwsze udane modele *deep learning* dla *lip-reading* łączyły dwie kluczowe składowe: konwolucyjną sieć neuronową (CNN) jako ekstraktor cech z obrazów oraz sieć rekurencyjną (RNN), najczęściej z komórkami długiej pamięci LSTM, do modelowania dynamiki sekwencji (Fenghour i in., 2021; Erbey, Barışçı, 2022; Lin i in., 2017). Taka architektura pozwalała uczyć się zarówno aspektów przestrzennych (układ ust), jak i czasowych (zmiany w czasie). Przykładem jest słynny model LipNet (Assael i in., 2017) – pierwsza w pełni *end-to-end* sieć do rozpoznawania całych zdań z ruchu ust. W LipNet zastosowano 3 warstwy CNN 2D (wyodrębniające cechy wizualne z kolejnych klatek) połączone z 2-warstwowym Bi-LSTM, a na wyjściu użyto mechanizmu CTC (*Connectionist Temporal Classification*) do generowania sekwencji znaków. Model ten, trenując się na dużym korpusie GRID (34 tys. zdań, prosty ograniczony słownik), osiągnął ~95% dokładności zdań, deklasując wcześniejsze metody (dla porównania – wynik doświadczonych ludzkich ekspertów to ~52%). Innym przykładem jest podejście Stafylakisa i Tzimiropoulosa (2017), które połączyło wydajną sieć ResNet-34 (do ekstrakcji cech z każdej klatki) z warstwami Bi-LSTM – osiągając wtedy rekordowe wyniki zarówno dla pojedynczych słów (97% na zbiorze GRID), jak i ciągłych ujęć *in the wild* (ponad 76% rozpoznawalności słów na 500-słownym zbiorze LRW). Architektury CNN+RNN stały się bazą wielu kolejnych prac; stosowano różne warianty: głębsze sieci konwolucyjne (np. VGG, Inception) jako *front-end* oraz sieci LSTM lub GRU (jednokierunkowe i dwukierunkowe) jako *back-end* (Fenghour i in., 2021; Erbey, Barışçı, 2022). Niektóre systemy zamiast przewidywać ciągi znaków dokonywały klasyfikacji do ustalonego słownika słów (przydatne np. do rozpoznawania komend). Ogólnie podejście CNN+RNN zdominowało dekadę 2015-2020 w dziedzinie *visual speech recognition*, przynosząc ogromny skok jakości.

Architektury 3D-CNN (cechy spatiotemporalne)

Alternatywą dla kombinacji 2D CNN + RNN stały się modele wykorzystujące konwolucje trójwymiarowe (3D CNN) lub ogólniej *spatiotemporal convolution*. Są to sieci, które uczą się filtrów konwolucyjnych działających jednocześnie w dwóch wymiarach przestrzennych (obrazu) i w wymiarze czasu – co pozwala od razu wydobywać lokalne cechy dynamiczne (ruch) bez osobnej warstwy rekurencyjnej (Ma, Sun, 2025; Wang i in., 2022). Przykładem jest model zaproponowany przez Zhanga i współautorów (2019, omówiony w Wang et al., 2022), w którym wykorzystano splot 3D do analizowania sekwencji obrazów ust, a następnie wyjście 3D-CNN sekwencyjnie przetwarzano już klasyczną siecią 2D-CNN (tzw. podwójna konwolucja: najpierw na rozdzielczości czasowej, potem przestrzennej). Modele z konwolucją czasową często cechują się większą równoległością przetwarzania niż RNN, przez co są łatwiejsze do trenowania na GPU. W literaturze opisano wiele wariantów – od prostych *Temporal Convolutional Networks (TCN)* po bardziej złożone architektury. Często stosuje się też połączenie: najpierw warstwy 3D-CNN, a potem klasyfikator sekwencyjny (np. TCN + CTC). Spatiotemporal CNN okazały się skuteczne zwłaszcza dla rozpoznawania pojedynczych słów i krótkich fraz. Przykładowo, Wang i in. (2022) zaproponowali model *Vision Transformer* z wbudowanym 3D-CNN, który na zbiorze LRW osiągnął ~89,5% dokładności – najlepszy wynik bez użycia dodatkowych danych. Zaletą 3D-CNN jest zdolność do jednoczesnego uchwycenia kształtu ust i jego zmian między sąsiadującymi klatkami (np. ruchu szczęki, pojawiania się zębów/przełyku itp.). Wyzwanie stanowi jednak większa złożoność obliczeniowa – filtry 3D mają więcej parametrów, co utrudnia trenowanie bardzo głębokich modeli na ograniczonych danych (tzw. ryzyko nadmiernego dopasowania). Dlatego często stosuje się wieloskładowe podejścia hybrydowe – np. łączenie bloków ResNet (uczących cech przestrzennych) z wstawkami 3D (dla dynamiki) lub stosowanie konwolucji separowalnych w czasie (np. model *Time-Depth Separable CNN*).

Mechanizmy uwagi i transformatory

Najnowsze lata przyniosły adaptację do zadania czytania z ust modeli opartych na mechanizmie *attention* oraz Transformerów, które zrewolucjonizowały

przetwarzanie języka naturalnego. Modele te rezygnują z recyrencji na rzecz mechanizmu *self-attention*, który potrafi globalnie modelować zależności w sekwencji. Już około 2019 roku pojawiły się pierwsze prace, w których moduły uwagi poprawiały dokładność rozpoznawania mowy wizualnej – np. poprzez selektywne skupianie się na kluczowych ramach (momentach artykulacji) lub ważnych obszarach obrazu (np. okolice ust kontra inne części twarzy) (Fenghour i in., 2021; Ma i Sun, 2025). W 2020-2021 zaczęto stosować pełne architektury Transformer, często łączone z CNN. Przykładowo, model *Vision Transformer + CTC* opisany przez Martineza i in. (omówiony w Wang i in., 2022) osiągnął bardzo dobre wyniki w rozpoznawaniu ciągłych zdań, zastępując LSTM blokiem Transformer. Zaletą uwagi jest umiejętność modelowania długoterminowych zależności – ważne np. dla uchwycenia kontekstu zdaniowego i uwzględnienia informacji językowej (składni, częstości słów). Nowoczesne systemy często wykorzystują tzw. architekturę *encoder-decoder Transformer*, analogiczną jak w tłumaczeniu maszynowym: enkoder (np. typu Conformer czy Swin Transformer) przetwarza sekwencję obrazów ust na wewnętrzną reprezentację, a dekoder Transformer generuje sekwencję wyjściową (np. liter lub słów) uwzględniając kontekst językowy. Renomowane grupy (np. z Baidu, DeepMind) zaprezentowały modele sekwencyjne z uwagą, które dla długich zdań potrafią zbliżyć się do skuteczności rozpoznawania audio. Należy wspomnieć, że mechanizm uwagi bywa też używany w połączeniu z CNN/RNN – np. moduły CBAM (*Convolutional Block Attention Module*) integrujące mapy uwagi kanałowej i przestrzennej z wyjściami CNN poprawiły wyniki w chińskim systemie *lipreadingu* Fu i Ni (2023). Ogółem, architektury z uwagą są obecnie jednym z głównych kierunków badań – umożliwiają bowiem bardziej efektywne wykorzystanie informacji o kontekście wypowiedzi niż klasyczne LSTM, choć wymagają bardzo dużych zbiorów danych do treningu ze względu na znaczną liczbę parametrów (Ma, Sun, 2025; Deshpande i in., 2025).

Inne podejścia głębokie

W literaturze znaleźć można także mniej konwencjonalne techniki. Jedna z nich to wykorzystanie sieci neuronowych (SNN) i przetwarzania impulsowego. Ponieważ ruch ust ma charakter sygnału czasowego, próbowano modelować go za pomocą sieci neuronowych przetwarzających dyskretne zdarzenia

(tzw. *spiking networks*). Zhang i wsp. (2021, przywoływani przez Ma, Sun, 2025) zaproponowali sieć kolczystą do rozpoznawania sylab, jednak jej skuteczność była niższa niż klasycznych podejść CNN/RNN. Ciekawą ścieżką jest natomiast użycie kamer zdarzeniowych (*event cameras*) do rejestrowania ruchu ust. Kamery te zapisują zmiany jasności w pikselach z wysoką częstotliwością (mikrosekundy) jako asynchroniczny strumień zdarzeń. Wstępne prace (Ma, Sun, 2025) sugerują, że takie sensory mogą uchwycić szybkie ruchy artykulacyjne niewidoczne w klasycznych nagraniach wideo, potencjalnie ułatwiając rozpoznawanie. Wymaga to jednak nowych algorytmów – często łączy się to z sieciami kolczystymi lub innymi architekturami dostosowanymi do danych zdarzeniowych.

Podsumowując, pod wpływem technik głębokiego uczenia nastąpiła znacząca zmiana paradygmatu w automatycznym czytaniu z ust. Tabela 1 zestawia cztery główne klasy podejść, omawiane powyżej, wraz z charakterystyką wykorzystywanych przez nie cech i modeli. W kolejnej części rozdziału przyjrzymy się, jakie konkretne wyniki udało się osiągnąć dzięki tym metodom oraz na jakich danych były one testowane.

Wyniki przeglądu badań

W celu syntetycznego porównania dotychczasowych osiągnięć badawczych zebrano informacje o wybranych modelach AI do rozpoznawania mowy z ruchu ust oraz o najczęściej wykorzystywanych zbiorach danych. Tabela 1 przedstawia porównanie reprezentatywnych podejść technicznych i ich skuteczności, zaś Tabela 2 podsumowuje właściwości kluczowych korpusów danych używanych w tej dziedzinie (Fenghour i in., 2021; Rahmatullah i in., 2025; Deshpande i in., 2025).

Jak wynika z Tabeli 1, na ograniczonych zadaniach (np. pojedyncze słowa z ustalonego słownika, zdania o prostej składni) współczesne modele AI osiągają bardzo wysokie wyniki – często powyżej 90%, a w niektórych przypadkach dorównują lub przewyższają umiejętności ludzkie (Fenghour i in., 2021). Dla przykładu, LipNet czy późniejsze modele ResNet+LSTM prawie bezbłędnie odczytują zdania z korpusu GRID, w którym składnia i słownictwo są mocno ograniczone. Jednak w miarę przechodzenia do trudniejszych, realnych danych (dłuższe wypowiedzi, dowolne słowa, różni mówcy i warunki) dokładność spada – na dużych korpusach typu LRS czy LRW-1000 najlepsze dotąd modele osiągają 40-50%

rozpoznawania (Rahmatullah i in., 2025; Wang i in., 2022), co pozostawia jeszcze duży margines do poprawy.

Tabela 1. Przegląd wybranych modeli do automatycznego czytania z ust z wykorzystaniem głębokiego uczenia. Zestawiono charakterystyki architektury oraz uzyskane wyniki na standardowych zbiorach danych

Model (rok)	**Główne cechy architektury**	**Zbiór danych (język)**	**Dokładność / Wynik**
LipNet – Assael i in. (2017)	end-to-end CNN + BiLSTM, CTC (rozpoznawanie zdań)	GRID (ang.; proste zdania)	95,2% dokładność zdań (WER ~5%) – przewyższa wynik ludzki
ResNet+LSTM – Stafylakis, Tzimiropoulos (2017)	Głęboki ResNet34 (ekstrakcja cech) + BiLSTM	GRID / LRW (ang.; słowa)	97,0% słowa na GRID; 76,2% słowa na LRW (500 słów)
Deep AVSR – Chung i in. (2017)	CNN + LSTM z połączeniem audiowizualnym (*Watch, Listen, Attend & Spell*)	LRS2 (ang.; zdania)	~50% wskaźnik BLEU (ok. 50% zdań poprawnie) – system audiowizualny (audio+video)
TMseq2seq – Afouras i in. (2018)	Transformer Encoder + Decoder seq2seq (uwaga wielogłowa)	LRS3 (ang.; zdania TED)	~41% WER (błąd słowa) w trybie wizualnym; ok. 2× lepszy niż wcześniejsze LSTM
Shillingford, i in (2019)	Masowy model *seq2seq* (DFCNN + LSTM, 2 tys. godz. danych)	własny zbiór (ang.; zdania)	40,9% WER dla zdań (setki godzin YouTube) – pierwszy model skalujący się na duże dane
Vision Transformer – Ma i in. (2022)	3DCNN + 2D SwinTransformer, klasyfikacja słów	LRW (ang.; 500 słów)	89,5% dokładność (Top1) – stan na 2022 r., bez danych dodatkowych (Wang i in., 2022)
EfficientGhostNet + GRU – Fu, Ni (2023)	Lekka sieć CNN (GhostNet zmodyfikowany) + BiGRU	własny (chiń.; 50 słów)	~85% dokładność słów – model zoptymalizowany pod urządzenia mobilne (Fu i in., 2023)

Model (rok)	Główne cechy architektury	Zbiór danych (język)	Dokładność / Wynik
SwinTransformer 3D + PTM – Zhang i in. (2023)	Model hybrydowy: SwinTransformer 3D + augmentacja *PartitionTime Masking*	LRW (ang.) / LRW 1000 (mandar.)	94,1% słowa na LRW; 45% słowa na LRW 1000 – najlepsze wyniki ok. 2023 r. (Wang i in., 2022)

Źródło: opracowanie własne na podstawie: Fenghour i in., 2021; Wang i in., 2022; Rahmatullah i in., 2025.

W Tabeli 2 zestawiono natomiast właściwości najpopularniejszych zbiorów danych wykorzystywanych w badaniach nad automatycznym czytaniem z ust. Są to zarówno korpusy kontrolowane (nagrania studyjne z ograniczonym słownikiem), jak i zbiory typu "*in the wild*" (klipy z telewizji, Internetu, o dużej różnorodności). Wielkość i złożoność tych zestawów danych w ogromnym stopniu determinują postęp – jak podkreślają Rahmatullah i in. (2025), brak dużych i zróżnicowanych danych przez lata hamował rozwój skutecznych modeli *lip-readingu*. Dopiero relatywnie niedawno – około 2016 r. – zaczęły powstawać naprawdę duże korpusy, co natychmiast przełożyło się na skok wyników (Fenghour i in., 2021; Rahmatullah i in., 2025).

Tabela 2. Przegląd wybranych publicznych zbiorów danych do wizualnego rozpoznawania mowy (lip-reading). Zestawiono charakterystykę korpusów: język, rodzaj i zakres wypowiedzi, liczebność i przykładowe osiągi najlepszych modeli (stan ok. 2023 r.)

Zbiór danych (rok)	Język / zakres mowy	Zawartość korpusu	Przykładowa skuteczność
GRID (2006)	angielski (proste zdania o ustalonej formie)	34 mówców, 1000 zdań każdy (ok. 33,5 tys. nagrań). Słownik ~51 słów, zdania 6-wyrazowe o stałej strukturze. Nagrania studyjne, frontalne wideo 25 Hz.	95% dokładność zdań (LipNet 2017). Ludzki poziom: ok. 50% (Assael i in., 2017).

Zbiór danych (rok)	Język / zakres mowy	Zawartość korpusu	Przykładowa skuteczność
CUAVE (2002)	angielski (ciągi cyfr)	36 mówców (19 par). Nagrania cyfr 0-9 (ciągi, izolowane). Ponadto nagrania w ruchu i z zakłóceniami. Łącznie ok. 7 tys. klipów.	~80-85% dla cyfr (modele z lat 2000-2010); rzadziej używany we współczesnych pracach.
AVLetters (1998) / AVLetters2 (2015)	angielski (alfabet)	AVLetters: 10 mówców, każda litera A–Z powtórzona 7 razy (łącznie 1820 klipów). AVLetters2: 5 widoków kątowych × 20 mówców × 26 liter (2600 klipów).	>90% dokładność liter w najnowszych modelach (Fenghour i in., 2021).
OuluVS (2009) / OuluVS2 (2015)	angielski (krótkie frazy)	OuluVS: 20 mówców, 10 fraz (np. “Excuse me”) – łącznie 200 nagrań. OuluVS2: 53 mówców, nagrania 5 słów z 5 kątów (0°, 30°, 45°, 60°, 90°) – ~2650 klipów.	>90% dokładność słów (stan ok. 2018); małe zbiory, głównie do testów *multi-view*.
LRW – Lip Reading in the Wild (2016)	angielski (słowa, *in the wild*)	~500 klas słów (brytyjski angielski, różne tematy). Nagrania z BBC. Ponad 1000 wystąpień na klasę, łącznie >500k klipów, długość ~1,16 s każdy. Duża zmienność mówców, ujęć, oświetlenia.	83% dokładność Top 1 (2018); ok. 94% (SOTA, 2023) (Wang et al.., 2022).
LRS2 – Lip Reading Sentences (2017)	angielski (zdania z TV)	Część *pre-train*: ~96 tys. zdań. Część *train-val*: 45 839 zdań (do 5 s) z 144 godzin programów BBC. Test: 1243 zdań. Słownik >40k słów.	~50% WER (DeepMind, 2017), spadek do ~32% WER (2021, Transformer; Rahmatullah i in., 2025).

Zbiór danych (rok)	Język / zakres mowy	Zawartość korpusu	Przykładowa skuteczność
LRS3 (TED) (2018)	angielski (zdania z TED)	Nagrania z TED/TEDx; ~118 516 segmentów zdań (do 30 s, średnio 4 s) z 409 wystąpień. ~475 godzin wideo.	~59% WER (2018, *seq2seq*) → ~33% WER (2022, ConformerTransformer; Rahmatullah i in., 2025).
LRW1000 (2019)	mandaryński (słowa, *in the wild*)	1000 klas słów (różne sylaby; każde słowo 1-2 znaki). 718 018 próbek wideo od >2000 osób (programy telewizyjne).	Tylko 43-45% dokładność Top 1 najlepszych modeli (2021-2023) – znacznie trudniejsza niż LRW (Wang i in., 2022).

Źródło: opracowanie własne.

Tabela 2 ilustruje rosnącą skalę i złożoność nowszych zbiorów danych. Wczesne korpusy (AVLetters, OuluVS) liczyły kilkaset nagrań i ograniczały się do pojedynczych słów czy liter, podczas gdy współczesne (LRW, LRS, LRW-1000) to setki tysięcy próbek reprezentujących realne warunki. Ten skok ilościowy był krytyczny – jak zauważają Rahmatullah i wsp. (2025), dopiero po 2014 r. pojawiło się >100 prac rocznie, gdyż badacze zyskali materiał do trenowania głębokich modeli. Mimo to nadal brakuje pełnej różnorodności – większość dużych zbiorów dotyczy języka angielskiego, co rodzi pytanie o przenoszalność metod na inne języki (np. tonalne, alfabet sylabiczny itp.) (Rahmatullah i in., 2025). Z tabeli widać też, że wyniki dla różnych korpusów bardzo się różnią. Najłatwiejsze (GRID – zamknięty kontekst) osiągnęły poziom niemal perfekcyjny, podczas gdy najbardziej wymagające (LRW-1000, spontaniczne zdania) pozostają dalekie od doskonałości. Ta rozbieżność jest istotna przy rozważaniu zastosowań praktycznych, co omówiono w następnym rozdziale.

Dyskusja

Zastosowania i znaczenie technologii

Automatyczne rozpoznawanie mowy z ruchu ust ma szereg potencjalnych zastosowań o dużym znaczeniu społecznym. Jednym z najważniejszych jest wsparcie komunikacyjne dla osób niesłyszących i niedosłyszących. Choć

implanty ślimakowe i aparaty słuchowe poprawiają percepcję dźwięku, wiele osób z ubytkiem słuchu i tak polega na czytaniu z ust, zwłaszcza w hałasie lub gdy rozmówca nie mówi wyraźnie (Battista i in., 2025; Santana i in., 2025). Systemy automatycznego czytania z ust mogłyby w przyszłości działać jako "tłumacz", zamieniając ruch ust rozmówcy na napisy w czasie rzeczywistym – analogicznie do napisów w filmach, ale generowanych na żywo. Badania Santany i in. (2025) pokazały, że czytanie z ust jest efektywną strategią poprawy rozumienia mowy przez osoby z ubytkiem słuchu, zwłaszcza w połączeniu z innymi metodami (np. aparatami słuchowymi). Automatyzacja tej strategii za pomocą AI może zatem znacząco zwiększyć dostępność informacji i jakość życia takich osób. Już teraz powstają prototypy aplikacji mobilnych, które potrafią rozpoznawać krótkie wypowiedzi z kamery smartfona i wyświetlać tekst – choć na razie ich skuteczność bywa ograniczona do prostych zdań i pojedynczych mówców (Erbey, Barışçı, 2022).

Drugim istotnym obszarem jest rozpoznawanie mowy w warunkach zakłóconego lub niedostępnego dźwięku. W wielu sytuacjach mikrofony zawodzą – np. z powodu głośnego szumu (przemysł, koncerty) albo gdy źródło dźwięku jest daleko. Człowiek do pewnego stopnia radzi sobie, patrząc na twarz mówiącego. Podobnie systemy audio-wizualnego rozpoznawania mowy (AV-ASR) integrujące sygnał dźwiękowy i obraz warg potrafią osiągnąć lepsze wyniki niż klasyczne ASR oparte wyłącznie na audio (Potamianos i in., 2003; Rahmatullah i in., 2025). Na przykład, gdy w nagraniu ktoś mówi w hałasie ulicznym – model audio-wizualny może wciąż poprawnie zidentyfikować słowa dzięki sygnałom wizualnym, podczas gdy model audio sam może się mylić. Ephrat i in. (2017, cyt. za Rahmatullah i in., 2025) pokazali system, który z niemego wideo potrafił częściowo odtworzyć przybliżony sygnał mowy (*Vid2Speech*), a Afouras i in. (2018) – demonstrację generowania mówiącej twarzy z samego audio (problem odwrotny, ale blisko spokrewniony). To wszystko wskazuje na komplementarność modalności audio i wideo. W praktyce przyszłe systemy rozpoznawania mowy w smartfonach czy robotach mogą korzystać z kamery do "wspomagania" rozpoznawania w trudnych warunkach akustycznych.

Innym polem jest bezpieczeństwo i nadzór. Czytanie z ust na nagraniach z monitoringu mogłoby pomóc zrozumieć, co mówią osoby obserwowane, nawet jeśli nie słychać dźwięku. W filmach sensacyjnych często pojawia się scena, gdzie

służby specjalne "czytają z ust" podejrzanych na odległość – AI mogłaby kiedyś to umożliwić na większą skalę. Rzecz jasna, wiąże się to z wątpliwościami natury prywatności i etyki (potencjalna masowa inwigilacja poprzez kamery CCTV), które należałoby rozważyć (Battista i in., 2025).

Kolejna kategoria to interfejsy człowiek-komputer w warunkach ciszy. Wyobraźmy sobie asystenta głosowego, z którego chcemy skorzystać w cichym biurze czy bibliotece – nie możemy głośno mówić, ale możemy szeptać lub poruszać ustami bezgłośnie. System *lip-reading* mógłby rozpoznać naszą wypowiedź i wykonać polecenie, nie emitując hałasu. Już teraz rozwijane są algorytmy rozpoznające tzw. *silent speech*, czyli mowę bezgłośną – czasem z pomocą specjalnych czujników (EMG mięśni twarzy, kamera termowizyjna itp.), a czasem tylko kamery. Przykładowo, praca Kumar i in. (2020, cyt. za Deshpande i in., 2025) wprowadziła system "Hush-Hush Speak" rekonstruujący mowę szeptaną z wideo za pomocą modelu generatywnego. Takie technologie mogą znaleźć zastosowanie w dyskretnym sterowaniu urządzeniami (np. wydawaniu poleceń smartfonowi bez mówienia na głos).

Warto też wspomnieć o wykorzystaniu czytania z ust w połączeniu z innymi modalnościami w systemach wielomodalnej komunikacji. Przykładowo, w rzeczywistości wirtualnej (VR) animowanie awatara postaci ludzkiej wymaga synchronizacji ruchu ust z mową, co można wspomóc algorytmami *lip-syncu*. W medycynie analiza wzorców ruchu ust może wesprzeć diagnozę zaburzeń mowy lub zaburzeń neurologicznych. Wreszcie technologia ta ma walor naukowy: pomaga badać proces integracji słuchowo-wizualnej (efekt McGurka – iluzja słuchowa pod wpływem bodźca wzrokowego) i lepiej rozumieć, jak ludzki mózg "łączy" obraz i dźwięk w percepcji mowy (Battista i in., 2025; Potamianos i in., 2003).

Ograniczenia obecnych metod

Pomimo imponujących postępów opisanych w rozdziałach poprzednich, istnieje wiele wyzwań i ograniczeń, z którymi mierzą się obecne systemy rozpoznawania mowy z ruchu ust. Wskazują je zarówno rezultaty licznych eksperymentów, jak i kompleksowe przeglądy (Rahmatullah i in., 2025; Santana i in., 2025; Deshpande i in., 2025). Poniżej omówiono kluczowe problemy.

Niedobór danych i generalizacja.

Trening głębokich sieci wymaga ogromnych zbiorów danych, a te – choć coraz większe – wciąż mogą być niewystarczające dla pełnej generalizacji. Większość obecnych modeli jest trenowana i testowana na względnie ograniczonych domenach (np. prezenterzy telewizyjni, wyraźna artykulacja). Gdy system zostanie użyty poza takim domenowym zbiorem – np. do rozpoznawania mowy przypadkowej osoby na ulicy – skuteczność może drastycznie spaść. Jest to problem niezależności od mówcy i warunków. Ludzie uczą się czytać z ust latami, oglądając wiele osób – analogicznie sieci potrzebują bardzo zróżnicowanych danych, by nie "przyuczyć się" do konkretnych twarzy, kształtów warg czy stylów mówienia. Jak zauważają Rahmatullah i in. (2025), nadal istnieją poważne wyzwania, takie jak niewystarczająca ilość danych, nieadekwatne warunki środowiskowe i różnorodność językowa, które ograniczają uniwersalność modeli. Przykładowo, model wytrenowany na brytyjskim angielskim może słabo radzić sobie z odczytywaniem amerykańskiego angielskiego (inne akcenty, ruchy ust), a tym bardziej z innym językiem – bo różnice fonetyczne i artykulacyjne są znaczne. Zatem brak wielojęzycznych, wielodialektalnych zbiorów to poważne ograniczenie.

Wizemy i kontekst językowy.

Jak już omawiano, wiele fonemów jest wizualnie nierozróżnialnych. Oznacza to, że system odczytujący mowę z ust ma fundamentalną niejednoznaczność, którą musi rozwiązać inaczej – poprzez kontekst językowy. Człowiek radzi sobie z wizemami dzięki temu, że wie, jakie słowa w danym języku są możliwe, więc może zgadnąć na podstawie części informacji. System AI musi czynić podobnie, integrując model języka (np. poprzez warstwę dekodera albo użycie statystyk n-gramów). Niemniej, w obecnych modelach – zwłaszcza pracujących na poziomie słów – czasem brakuje głębszego rozumienia kontekstu. Model może popełniać błędy semantycznie niedorzeczne, np. odczytać zdanie jako pozbawione sensu, bo wybrał wizualnie pasujące słowo, które jednak nie pasuje do reszty zdania. Problem wizemów szczególnie utrudnia rozpoznawanie krótkich wypowiedzi i pojedynczych słów bez kontekstu – np. rozróżnienie "mat" vs "bat" tylko z obrazu ust jest prawie niemożliwe (Bear, Harvey, 2017; Rahmatullah i in., 2025).

Warunki oświetleniowe, pozycja kamery i jakość wideo.

Większość badań wykorzystuje dość dobrej jakościowo dane – frontalne ujęcia twarzy, przyzwoita rozdzielczość (≥ 100 px szerokość ust), dobre oświetlenie. W praktyce kamery monitoringu czy smartfony nie zawsze dają taki obraz. Silne cienie na twarzy, słabe światło, duży kąt (profil zamiast *en face*), niska rozdzielczość – to wszystko obniża skuteczność czytania z ust. Badania wskazują, że przy rozdzielczości ust < 20 px systemy praktycznie tracą zdolność rozpoznawania (Deshpande i in., 2025). Również opóźnienia w klatkażu (np. kamera 5 fps) sprawiają, że ruch wydaje się skokowy, co utrudnia analizę. Niektóre prace starają się radzić z tym poprzez *superresolution* lub poprawę jakości obrazu jako etap wstępny, ale to wciąż aktywna dziedzina (Ma, Sun, 2025).

Złożoność obliczeniowa i opóźnienia.

Głębokie modele *lip-reading* bywają bardzo duże (np. transformatory z setkami milionów parametrów) i wymagają potężnych GPU do działania w czasie rzeczywistym. To ogranicza ich zastosowanie np. w urządzeniach mobilnych. Powoli pojawiają się próby optymalizacji – np. Fu i Ni (2023) zaproponowali lekką sieć Efficient GhostNet, redukując liczbę parametrów bez dużej utraty dokładności. Mimo to osiągnięcie płynnego działania, np. na smartfonie z samą kamerą (bez wysyłania danych do chmury), nadal jest wyzwaniem. Trzeba też pamiętać o opóźnieniu predykcji: przy rozpoznawaniu zdań modele często potrzebują "obejrzeć" całą sekwencję, zanim wydadzą wynik (zwłaszcza te z dekoderem *seq2seq*). To oznacza, że przy długiej wypowiedzi system zadziała dopiero po jej zakończeniu. Dla interaktywnych zastosowań lepsze są modele *streamingowe* lub działające w oknach czasowych. CTC oferuje tu przewagę, bo pozwala częściowo *online* dekodować, ale większość badań nadal koncentruje się na trybie *offline*.

Bezpieczeństwo i etyka.

Zastosowania nadzorcze tej technologii niosą pytania etyczne: czy wolno maszynowo odczytywać z ust ludzi bez ich wiedzy? To budzi obawy o prywatność – ktoś może być de facto "podsłuchiwany", nawet jeśli nie rozmawia przez telefon, a jedynie rozmawia twarzą w twarz, ale w zasięgu kamery. Już teraz

technologie typu *lip-reading CCTV* budzą kontrowersje (Battista i in., 2025). Z technicznego punktu widzenia ograniczeniem (a z etycznego może zaletą) jest tu to, że systemy te wciąż są zawodne w warunkach rzeczywistych – więc raczej nie są masowo stosowane. Jednak przy dalszym postępie prawdopodobnie pojawi się konieczność regulacji prawnych dotyczących wykorzystania tej technologii.

Kierunki dalszych badań

Biorąc pod uwagę wymienione ograniczenia, społeczność naukowa wyznaczyła szereg kierunków badań, które mają na celu udoskonalenie automatycznego rozpoznawania mowy z ruchu ust (Fenghour i in., 2021; Rahmatullah i in., 2025; Deshpande i in., 2025).

Zwiększanie i urozmaicanie danych treningowych

Należy tworzyć większe i bardziej zróżnicowane zestawy danych – obejmujące różne języki, dialekty, grupy wiekowe, warunki nagrania. Szczególnie paląca jest potrzeba korpusów dla języków innych niż angielski (np. chiński LRW-1000 jest krokiem w tę stronę, ale przydałby się np. duży zbiór dla języków z innym systemem fonetycznym – arabskiego, hindi, hiszpańskiego, polskiego). Ponadto dane w warunkach naturalnych – nagrania z różnych kątów, w ruchu – pomogą modelom lepiej generalizować. Jedną z dróg jest generowanie danych syntetycznych – np. użycie modeli generatywnych do tworzenia nowych nagrań ust mówiących zadane frazy (speech-to-video), co może służyć augmentacji danych dla rzadkich słów czy nietypowych warunków (Deshpande i in., 2025).

Uczenie niewspierane i transfer wiedzy

Aby zmniejszyć zależność od etykietowanych danych (transkrypcji), intensywnie bada się metody *self-supervised learning*. Modele mogłyby uczyć się reprezentacji ruchu ust bez transkrypcji, np. predykując przyszłe klatki, dopasowując zdania do klipów (*contrastive learning*) itp. (Ma, Sun, 2025). Taki wstępnie wytrenowany model można potem dostroić na mniejszej liczbie danych oznaczonych. Innym podejściem jest *transfer learning* z pokrewnych zadań: np. model do

rozpoznawania mowy audio lub do rozpoznawania mimiki twarzy może dostarczyć część reprezentacji dla modelu *lip-reading* (Fenghour i in., 2021).

Integracja z modelem językowym

W miarę postępów w NLP, rozsądnym kierunkiem jest ściślejsze powiązanie systemów czytania z ust z nowoczesnymi modelami językowymi. Obecne architektury *seq2seq* już to robią w pewnym stopniu, ale można pójść dalej: np. użycie dużego modelu językowego do autokorekty wyjścia systemu *lip-reading*. Model mógłby wygenerować listę najlepszych hipotez, a LLM ocenić, która jest najbardziej sensowna w kontekście (Deshpande i in., 2025). Integracja na poziomie semantyki mogłaby częściowo rozwiązać problemy związane z wizemami i błędami gramatycznymi.

Multimodalność – systemy audiowizualne

Badania pokazują, że modele uczone jednocześnie na audio i wideo często uczą się lepszych reprezentacji – nawet jeśli potem audio nie jest dostępne (Potamianos i in., 2003; Rahmatullah i in., 2025). Dlatego warto kontynuować rozwój modeli multimodalnych. W szczególności technika *teacher-student*: model audio (łatwiejszy do trenowania, bo ASR audio jest dojrzały) naucza model wizualny, przekazując mu wskazówki (*knowledge distillation*). Integracja audio-wideo jest celem samym w sobie – przyszłe systemy rozpoznawania mowy w realnych zastosowaniach powinny wykorzystywać wszystko, co dostępne (mikrofony + kamery), by zminimalizować błędy. Można też rozszerzyć modalność: np. łączyć czytanie z ust z rozpoznawaniem mowy migowej (ruch rąk) dla komunikacji z osobami głuchymi.

Nowe architektury sieciowe

Choć transformatory dominują, eksplorowane są nowe architektury mogące lepiej pasować do sekwencji wizyjnych – np. wspomniane sieci kolczyste (*spiking neural nets*) potencjalnie bardzo wydajne energetycznie, działające w czasie rzeczywistym na sprzęcie neuromorficznym (Ma, Sun, 2025). Jeśli uda się opracować SNN skutecznie rozpoznającą wzorce ruchu ust, mogłoby to otworzyć drogę

do miniaturowych układów (np. implantów rozumiejących mowę z mikrokamer). Innym trendem są modele z atencją czasowo-kanoniczną, potrafiące radzić sobie z różnymi tempami mowy.

Lepsze zrozumienie i interpretowalność

Większość modeli to tzw. czarne skrzynki – wiemy, że działają (lub nie), ale trudno powiedzieć, dlaczego np. pomyliły "mat" z "bat". Przyszłe badania mogą skupić się na uczynieniu modeli bardziej interpretowalnymi. Np. wizualizacje map uwagi pozwalają zobaczyć, na które klatki czy obszary model patrzył, podejmując decyzję (Deshpande i in., 2025). W kontekście prawnym (np. wykorzystania w sądzie odczytu maszynowego z monitoringu) potrzebna będzie ocena wiarygodności – system musiałby sygnalizować, na ile pewne jest rozpoznanie. Kalibracja takich miar zaufania dla mowy wizualnej jest osobnym wyzwaniem.

Podsumowując, perspektywy rozwoju automatycznego czytania z ust są bardzo obiecujące. Jak zauważają Rahmatullah i in. (2025), postęp w ostatniej dekadzie był ogromny – nowe techniki głębokiego uczenia pozwoliły osiągnąć przekonujące wyniki *state-of-the-art*, choć wciąż pozostaje wiele do zrobienia, aby zrównać wydajność systemów wizualnych z audio i pokonać takie przeszkody, jak niedobór danych czy niejednoznaczność wizualna. Trend wskazuje, że integracja modalności, większe modele i lepsze algorytmy uczenia będą sukcesywnie zmniejszać tę lukę.

Wnioski

Rozpoznawanie mowy na podstawie ruchu ust i warg jest zagadnieniem o długoletniej historii badawczej, które dzięki połączeniu z nowoczesnymi technikami sztucznej inteligencji wkroczyło w nową fazę rozwoju. W niniejszym przeglądzie dokonano podsumowania stanu badań w tym zakresie. Wprowadzenie nakreśliło znaczenie problemu – zarówno dla osób z ubytkiem słuchu, jak i w kontekście wzbogacania klasycznego rozpoznawania mowy w hałasie czy budowy multimodalnych interfejsów. Przypomniano też wyzwania, takie jak niejednoznaczność wizualna głosek (problem wizemów), które utrudniają to zadanie.

W części metod prześledzono ewolucję technik: od wczesnych podejść opartych na cechach geometrycznych i HMM, poprzez pojawienie się sieci neuronowych CNN+RNN zdolnych do automatycznej ekstrakcji cech z obrazów ust, aż po współczesne architektury Transformer i hybrydowe modele *spatiotemporalne*. Na Rysunku 1 zobrazowano taksonomię tych metod, uwzględniając różne kategorie cech (wizualne vs. czasowo-przestrzenne), mechanizmy uwagi oraz nowatorskie podejścia (sieci kolczyste, kamery zdarzeniowe). Wyniki przeglądu zestawione w tabelach pokazują, że na ograniczonych zadaniach (np. zamknięty słownik słów) systemy AI osiągają już znakomite rezultaty (rzędu 90-95% poprawności), natomiast w warunkach zbliżonych do rzeczywistych wciąż istnieje spora luka do zapełnienia – dokładności rzędu 40-50% wskazują, że jest miejsce na poprawę (Rahmatullah i in., 2025; Wang i in., 2022). Analiza najczęściej używanych korpusów danych (Tabela 2) ujawnia dramatyczny wzrost skali dostępnych danych w ostatnich latach, co było kluczowe dla sukcesu metod głębokiego uczenia.

W dyskusji podkreślono praktyczne zastosowania technologii czytania z ust (wspieranie niesłyszących, rozpoznawanie mowy w hałasie, nadzór i bezpieczeństwo, interfejsy cichej mowy), ale także bariery, które utrudniają ich pełną realizację. Obecne ograniczenia – takie jak niedostatek danych spoza języka angielskiego, wrażliwość na warunki nagrania, duża złożoność obliczeniowa modeli czy kwestie prywatności – wymagają dalszych badań. Zaproponowano szereg kierunków przyszłego rozwoju: od gromadzenia nowych wielojęzycznych zbiorów danych, przez uczenie samonadzorowane i transfer wiedzy z modeli audio, po udoskonalenia architektur (np. integracja z modelami językowymi, zastosowanie sieci Transformer również po stronie ekstrakcji cech, rozwój modeli multimodalnych audio-wizualnych). Zwrócono również uwagę na aspekt etyczny i potrzebę interpretowalności – tak, by rosnące możliwości technologiczne szły w parze z zaufaniem do systemów AI.

Reasumując, rozpoznawanie mowy z ruchu ust przeszło drogę od ciekawostki naukowej do poważnej dziedziny badań nad sztuczną inteligencją. Wykorzystanie głębokich sieci neuronowych przyniosło ogromny skok jakościowy – do poziomu, na którym systemy te mogą realnie konkurować z ludzkimi umiejętnościami w ograniczonych warunkach (Fenghour i in., 2021). Jednak dopiero rozwiązanie otwartych problemów (wizemy, generalizacja, dane) pozwoli na

powszechne wdrożenia. Wysiłki badawcze trwają: jak podaje najnowszy systematyczny przegląd literatury, w latach 2014-2024 opublikowano ponad 100 oryginalnych prac w tej tematyce, a tempo badań rośnie (Rahmatullah i in., 2025). Można oczekiwać, że w kolejnych latach będziemy świadkami dalszego zbliżania wydajności systemów czytania z ust do rozpoznawania mowy z dźwięku – być może aż do punktu, gdy obie te modalności połączą się, dając w pełni multimodalne systemy rozumienia mowy działające niezawodnie w różnych warunkach.

Literatura

1. Battista, M., Collesei, F., Orzan, E., Fantoni, M., Bottari, D. (2025). LipReading: Advances and Unresolved Questions in a Key Communication Skill. *Audiology Research, 15*(4), 89. https://doi.org/10.3390/audiolres15040089
2. Bear, H. L., Harvey, R. (2017). Phoneme-to-viseme mappings: the good, the bad, and the ugly. *Speech Communication*, 95, 40-67. https://doi.org/10.1016/j.specom.2017.07.001
3. Deshpande, S., Shirsath, K., Pashte, A., Loya, P., Shingade, S., Sambhe, V. (2025). A Comprehensive Survey of Advancement in Lip Reading Models: Techniques and Future Directions. *IET Image Processing, 19*(1). https://doi.org/10.1049/ipr2.70095
4. Erbey, A., Barışçı, N. (2022). A Survey on LipReading with Deep Learning. *Uluslararası Mühendislik Araştırma ve Geliştirme Dergisi (International Journal of Engineering Research and Development), 14*(2), 844-860. https://doi.org/10.29137/umagd.1038899
5. Fenghour, S., Chen, D., Guo, K., Li, B., Xiao, P. (2021). Deep LearningBased Automated LipReading: A Survey. *IEEE Access, 9*, 121184-121205. https://doi.org/10.1109/ACCESS.2021.3107946
6. Fu, Y., Lu, Y., Ni, R. (2023). Chinese LipReading Research Based on ShuffleNet and CBAM. *Applied Sciences, 13*(2), 1106. https://doi.org/ 10.3390/app13021106
7. Lin, B. S., Yao, Y. H., Liu, C. F., Lien, C. F., Lin, B. S. (2017). Development of Novel LipReading Recognition Algorithm. *IEEE Access, 5*, 794-801. https://doi.org/10.1109/ACCESS.2017.2649838
8. Ma, Y., Sun, X. (2025). Spatiotemporal Feature Enhancement for LipReading: A Survey. *Applied Sciences, 15*(8), 4142. https://doi.org/10.3390/app15084142

9. Mase, K. and Pentland, A. (1991), Automatic lipreading by optical-flow analysis. Syst. Comp. Jpn., 22: 67-76. https://doi.org/10.1002/scj.4690220607
10. Matthews, I., Cootes, T. F., Bangham, J. A., Cox, S., Harvey, R. (2002). Extraction of visual features for lipreading. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 24*(2), 198-213. https://doi.org/10.1109/34.982900
11. Miled, M., Messaoud, M. A. B., Bouzid, A. (2023). Lip reading of words with lip segmentation and deep learning. Multimedia Tools and Applications, *82*, 551-571. https://doi.org/10.1007/s11042-022-13321-0
12. Potamianos, G., Neti, C., Gravier, G., Garg, A., Senior, A. (2003). Recent advances in the automatic recognition of audiovisual speech. *Proceedings of the IEEE, 91*(9), 1306-1326. https://doi.org/10.1109/JPROC.2003.817150
13. Rahmatullah, G. M., Ruan, S.J., Wisesa, I. W. W., Li, L. P.H. (2025). Enhancing visual speech perception through deep automatic lipreading: A systematic review. *Computers in Biology and Medicine, 190*, 110019. https://doi.org/10.1016/j.compbiomed.2025.110019
14. Santana, R. S., Menezes, E. C., Ralin, V., Givigi, S. (2025). Lipreading as a communication strategy to enhance speech recognition in individuals with hearing impairment: A scoping review. *Disability and Rehabilitation: Assistive Technology, 20*(5), 1235-1246. https://doi.org/10.1080/17483107.2025.2449984
15. Wang, H., Pu, G., Chen, T. (2022). A lip-reading method based on a 3D convolutional vision transformer. *IEEE Access, 10*, 77205-77212. https://doi.org/10.1109/ACCESS.2022.3193231