Abstract

Visual Speech Recognition (VSR), or lipreading, is a significant branch of artificial intelligence research that uses visual cues to understand speech. It serves as an important augmentative communication tool when audio is unavailable or unreliable and forms the cornerstone of accessibility technologies for individuals with hearing impairments. Recent advances in deep learning, particularly through 3D Convolutional Neural Networks (3D-CNNs), Recurrent Neural Networks (RNNs) such as LSTM/GRU, and Transformer-based frameworks, have significantly increased the capacity to model the spatial and temporal dynamics of lip movements. However, because existing models and datasets generally focus on English, progress in languages with diverse features remains limited. Turkish is an underrepresented language that presents additional modeling challenges due to its agglutinative morphology, vowel harmony rules, and unique phoneme-viseme mappings. This study proposes a Turkish lipreading system incorporating modern deep learning methods specifically designed for the grammatical structure of Turkish to advance multilingual VSR research and support accessibility applications.

Öz

Görsel Konuşma Tanıma (VSR) ya da dudak okuma, konuşmaları anlamak için görsel işaretleri kullanan yapay zeka araştırmasının önemli bir dalıdır. Sesin mevcut olmadığı ya da güvenilmez olduğu durumlarda önemli bir destekleyici iletişim aracı olarak işlev görür ve işitme engelli bireyler için erişilebilirlik teknolojilerinin temel taşını oluşturur. Son zamanlarda derin öğrenmedeki gelişmeler, özellikle 3D Evrişimli Sinir Ağları (3D-CNN'ler), LSTM/GRU gibi Tekrarlayan Sinir Ağları (RNN'ler) ve Transformer tabanlı yapılar sayesinde, dudak hareketlerinin mekansal ve zamansal dinamiklerini modelleme kapasitesini büyük ölçüde artırmıştır. Ancak mevcut modeller ve veri setleri genellikle İngilizceye odaklandığı için, farklı özelliklere sahip dillerde ilerleme kısıtlı kalmaktadır. Türkçe, eklemeli morfolojik yapısı, sesli harf uyumu kuralları ve özgün fonem-vizem eşleşmeleri nedeniyle ek modelleme zorlukları sunan ve yeterince temsil edilmeyen bir dildir. Bu çalışma, çok dilli VSR araştırmalarını ilerletmek ve erişilebilirlik uygulamalarını desteklemek amacıyla, Türkçenin dilbilgisel yapısına özel olarak tasarlanmış modern derin öğrenme yöntemlerini içeren bir Türkçe dudak okuma sistemi önermektedir.

1. Introduction

Visual Speech Recognition (VSR) is a technology that enables understanding spoken language by observing only lip and facial movements [1, 8]. In recent years, it has become an important research topic in the fields of artificial intelligence and speech processing [8]. Lip reading offers an alternative means of communication, especially in situations where sound is inaudible or unreliable [7], and plays an important role in technologies developed for hearing-impaired individuals [1, 6].

Advances in deep learning—3D Convolutional Neural Networks (3D-CNN), recurrent models such as LSTM/GRU, and Transformer-based structures—have enabled more accurate modeling of lip movements both temporally and spatially [2, 3, 4, 9]. However, current datasets and models are mostly focused on English. This makes development difficult for languages with different phonetic structures [4, 8].

Turkish is one of the languages underrepresented in VSR studies. Its agglutinative structure, vowel harmony, and unique phoneme–viseme relationships create additional challenges for lip reading models[10]. Furthermore, the scarcity of Turkish lip reading datasets limits research in this field[11].

This study aims to propose a Turkish lip-reading system using modern deep learning methods suitable for the linguistic characteristics of Turkish. Thus, it is intended to contribute to accessibility applications and support multilingual VSR research.

2. Lip Reading

Lip reading relies solely on visual information obtained from lip and mouth movements [1, 5, 6, 7, 8]. Therefore, it is a more limited task compared to speech recognition based on sound [5, 7]. Nevertheless, VSR stands out as an important complementary method, especially in noisy or acoustically complex environments [5, 8]. Recent studies show that deep learning models can effectively capture the detailed spatial and temporal features of lip movements [2, 3, 4, 8, 9].

2.1 What is Lip Reading?

Lip reading is the process of automatically interpreting speech using visual features extracted from the lips, tongue, and surrounding facial areas [1, 7, 8]. This task requires modeling both the instantaneous positions in the image and the changes over time simultaneously [3, 4, 8, 9]. This difficulty is frequently mentioned in the VSR literature. End-to-end models such as LipNet, LRS2, and LRS3 have provided strong reference points and highlighted the importance of spatio-temporal integrated modeling [2, 3, 4, 8].

2.2 The Importance of Visual Speech Recognition

1. Noise Resistance:

   Visual cues enhance recognition performance by providing additional information when sound quality is poor [5, 8]. Therefore, VSR is an important part of multimedia systems

2. Accessibility:

   VSR technologies provide important support to hearing-impaired individuals and can be widely used in healthcare, education, and assistive communication systems [1, 6].

3. Silent and Secure Communication:

   In environments where voice recording is not possible or preferred, lip reading can be used as an effective silent communication method [7, 6].

2.3 Challenges in Lip Reading

2.3.1 Speaker Variability

Individual differences such as facial structure, speaking style, and speaking speed reduce the generalization ability of VSR models. Research clearly demonstrates that speaker-independent models generally perform worse and require large, diverse datasets [12].

2.3.2 Environmental Factors

Light changes, face occlusion, camera angle, and video quality affect the visibility of the lip region, making feature extraction difficult. Therefore, robust preprocessing steps and comprehensive data augmentation techniques are commonly used [12].

2.3.3 Vowel Ambiguity

Many phonemes are visually very similar (e.g., /p-b-m/). This limits the discrimination power of models relying solely on visual information. To mitigate this, time-series-based modeling and language model integration are frequently used methods [13].

3. Visual Speech Recognition Techniques

Visual speech recognition (VSR) systems for lip reading typically differ by the *type* of visual signal they extract and how they feed that signal to temporal models.

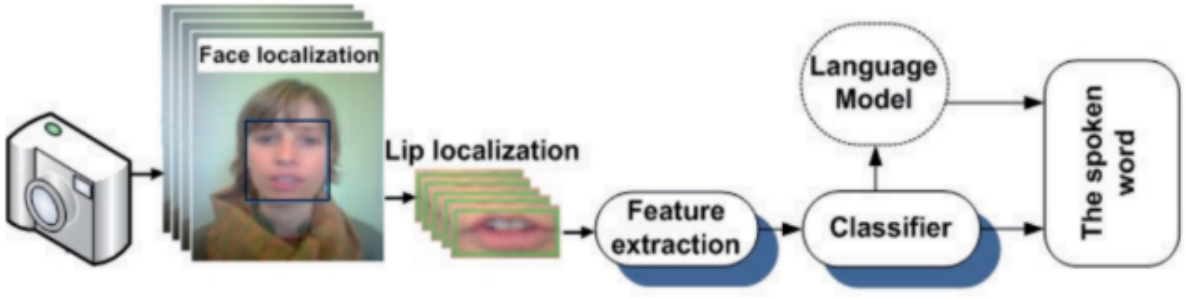

Figure 1. Visual Speech Recognition pipeline, adapted from [1].

3.1 Feature Extraction Methods

In VSR systems, the choice of feature extraction strongly affects recognition accuracy, robustness, and computational cost. There are two kinds of extraction methods:

1. Traditional Methods
    - Appearance-Based Features
    - Geometric-Based Features
2. Deep Learning Based Methods
    - Convolutional Neural Networks (CNNs)
    - Recurrent Neural Networks (RNNs)
    - Transfer Learning and Pre-trained Models

3.1.1 Traditional Methods

Features relying on visual appearance utilize the raw pixel values or mathematically transformed versions of the mouth area. They are generally simple to implement and excel at recording the full spectrum of visual cues.

Appearance-Based Features:

- **Pixel Intensity Features:** Uses the basic color and brightness of the lip region. It's simple but highly sensitive to noise and poor lighting. Techniques like PCA are often used to clean and reduce the data.
- **Discrete Cosine Transform (DCT):** Converts the image into frequency parts. It keeps only the low-frequency parts that define the overall shape and structure, helping to filter out unwanted noise.
- **Gabor Filters:** Uses multiple filters to capture detailed texture and edges at various angles and sizes. It is very effective at spotting subtle, dynamic changes in the lip area.
- **Local Binary Patterns (LBP):** An efficient way to describe the texture of the lips by comparing neighboring pixels. Because it is unaffected by general changes in brightness, it is robust against variations in illumination.

Geometric- Based Features:

- **Active Appearance Models (AAMs):** These are statistical models that combine both the shape (geometry) and texture (appearance) of the mouth. They are used to accurately track and extract these two types of features together.
- **Landmark-Based Features:** Features are defined by identifying specific points (landmarks) on the face or mouth, such as the corners. The distances and angles between these points are used as the primary measurements.These methods are preferred when **efficiency and stability** matter more than fine-grained detail.

3.1.2 Deep Learning Based Methods

Convolutional Neural Networks (CNNs): They can extract spatial features (even temporal features via 3D CNN models) from the input video frames. By applying convolutional filters, the network captures local patterns such as the curvature of the lips, the visibility of the tongue, and teeth positioning.

Recurrent Neural Networks (RNNs): They are able to process sequential data and they have memory to keep the previous time steps, which is essential for modeling the temporal evolution of speech.

Transfer Learning and Pre-trained Models: Training models from scratch require massive labeled datasets. These models can significantly reduce computational costs such as resources, training time and data, also they can accelerate the training process. [14]

Hybrid Approaches

Many recent works combine landmarks, appearance, and motion to exploit complementary strengths. This fusion improves robustness across speakers and recording conditions, though at the cost of a more complex pipeline.

3.2 Model Architectures

In VSR systems, the choice of model architecture is crucial for capturing the complex spatiotemporal characteristics of lip reading. There are various deep learning architectures, for instance Convolutional Neural Networks, Recurrent Neural Networks, Transformer-based models, and hybrid models. Each architecture has distinct advantages: CNNs are effective at extracting visual features, RNNs resolve temporal sequentiality, Transformers address long-range contextual dependencies, hybrid architectures that combine these approaches have demonstrated superior performance by leveraging the benefits of multiple models simultaneously.

3.2.1 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are usually used to extract visual features from the detected lip region. Although 2D CNNs are computationally efficient, they are limited to extracting spatial features and cannot capture the temporal dynamics, which are essential for continuous lip movements. Research conducted by Stafylakis and Tzimiropoulos [15] shows that treating frames in isolation fails to capture the short-term temporal dynamics, which they initially utilized 2D CNNs to process frames. In contrast, 3D CNNs can process both the temporal and spatial dimensions. Therefore, they can capture dynamic patterns of visual speech. [16]

3.2.2 Recurrent Neural Networks (RNN / LSTM / GRU)

RNNs model temporal dependencies in sequences of lip movements. Each frame represents a time step, and the RNN learns the progression of shapes/motions over time.

Variants:

- **LSTM (Long Short-Term Memory):** Captures long-range dependencies and reduces vanishing gradient problems.
- **GRU (Gated Recurrent Unit):** GRU is similar to LSTM but with fewer parameters and faster training.

3.2.3 Transformer-Based Models

Transformers handle long-term temporal dependencies efficiently using attention mechanisms. They can model global context better than RNNs. It is parallelizable, better at modeling longer sequences and robusting noise.

Pipeline:

1. CNN extracts per-frame features.
2. Features are fed to a Transformer encoder (sometimes encoder-decoder for sequence-to-sequence tasks).
3. Output predicts phonemes, words, or sentences.

3.2.4 Hybrid Deep Learning Architectures

Concept of Hybrid Deep Learning Architectures is that it combines CNNs for spatial features and RNNs/Transformers for temporal modeling. Some models also integrate attention mechanisms or 3D CNNs for spatiotemporal features. For example:

- CNN + LSTM: CNN extracts frame features, LSTM models time sequence.

- 3D CNN + Transformer: 3D CNNs capture both space and short-term temporal dynamics; Transformers capture longer-term dependencies.

We are going to use a hybrid model to increase the performance of our model.

4. Turkish Lip Reading

4.1 Characteristics of Turkish Language Affecting Lip Reading

The development of Automatic Lip Reading (ALR) systems for Turkish faces serious challenges stemming from the unique structure of the language. The fundamental problem is Visual Ambiguity (Viseme Ambiguity), caused by the Vowel Harmony rule, which complicates visual discrimination (e.g., the similarity between words like “güzel” and “kolay”). This visual ambiguity is further compounded by Turkish's Rich Morphology; while producing long and variable sequences derived from a single root, phonological structures such as liaison create transitions that can alter meaning but are difficult to perceive visually. Furthermore, the natural dynamics of the language and variations in pronunciation (e.g., “gardaş” instead of “kardeş,” ‘gidiyom’ instead of “gidiyorum”) weaken the model's ability to generalize.

Furthermore, minimal or absent lip movement in words like “iyi” and throat sounds reduces the reliability of visual data, driving systems toward models that are more adept at tracking temporal sequences. To overcome these challenges, current research focuses on datasets that capture natural variation (dialect, elision, stress) and include participants from different regions, rather than controlled recordings. However, given the general scarcity of data in the field, the number of studies that have generated lip-reading data in Turkish is quite limited. Among the analyzed studies, it is observed that Hadi Pourmousa & Üstün Özen[17] and Hasan Atila & Ali Sabaz[18] created their own original datasets, while Talya Tümer Sivri, Ali Berkol, and Hamit Erdem[19] worked on visual Turkish data. This situation reveals that the biggest obstacle in developing ALR systems for Turkish is still the lack of publicly available and extensive data sets.

4.1.1 Vowel Harmony

The Turkish Sound Harmony rule is the fundamental phonological mechanism that causes Visual Ambiguity (Viseme Ambiguity), the most critical challenge for Automatic Lip Reading (ALR) systems. This rule makes the lip and mouth positions of words with different phonemes visually indistinguishable. In the work of Hadi Pourmousa and Üstün Özen, the model's confusion of lip movements for adjectives such as “güzel” and “kolay” due to their excessive similarity is concrete evidence of this visual ambiguity. Talya Tümer Sivri, Ali Berkol, and Hamit Erdem also confirmed in their work that distinguishing similar lip movements is the biggest problem when working with visual data alone. As a result of this visual ambiguity, models based on single frames are insufficient, The fact that sentence recognition success (88.55%) in the work of Hasan Atila and Ali Sabaz is significantly higher than word recognition success (84.5%) proves that the difficulty arising from Sound Assimilations can only be overcome by using temporal context (temporal information). This situation directs ALR systems towards temporal models such as BGRU and Bi-LSTM.

4.1.2 Rich Morphology

The agglutinative and rich morphology structure of Turkish poses one of the greatest dynamic challenges that Automatic Lip Reading (ALR) systems must overcome. This structure allows for the derivation of numerous inflected forms from a single word root, producing longer and more variable lip sequences compared to standard languages, thereby increasing the length and complexity of the visual sequences the model must learn.

The work of Talya Tümer Sivri and others has indicated that word or sentence length is strongly correlated with the number of syllables and that speech rate varies significantly between speakers, demonstrating that temporal variation arising from morphological structure is a critical issue. When morphological complexity combines with the dynamics of everyday speech, it leads to different pronunciation forms, as observed by Hadi Pourmousa and Üstün Özen:

- **Verb Abbreviations:** Using “gidiyom” instead of “gidiyorum” (I'm going).

- **Regional Dialects:** Using “gardaş” instead of “kardeş” (brother).

Furthermore, phonological events such as liaison, as mentioned by Hasan Atila and Ali Sabaz, involve the linking of word-final consonants to the following vowel:

- **Liaison:** As in the examples “top aldı” and “topaldı,” it creates visual transitions that can change the meaning but are difficult to discern in lip movement.

As a result, Rich Morphology requires ALR systems to correctly classify not only a single word, but also a highly variable temporal flow.

4.1.3 Viseme Challenges Specific to Turkish

Some studies conducted on Turkish, such as “Lip Reading Using CNN for Turkish Numbers,” have used CNN-based single-frame models on limited word sets. However, these studies have not taken into account viseme similarities caused by phonological rules in Turkish, such as vowel harmony and vowel narrowing. The limited success of single-frame CNN models even in a narrow word domain such as number names indirectly demonstrates how difficult it is phonologically to distinguish lip shapes in Turkish and that viseme ambiguity is a structural problem. Therefore, it is to be expected that image-based CNN architectures alone will be insufficient for tasks requiring viseme discrimination in Turkish.

Viseme ambiguity in Turkish is a structural problem stemming from the phonological structure of the language, fundamentally limiting the ability of Automatic Lip Reading (ALR) systems to make definitive distinctions using visual information. An analysis highlighted that in some Turkish words (e.g., “iyi,” “40,” and “2”), the near-absence of lip movement leaves the system with zero visual clues for classification. This is the most concrete example of how the phonological structure weakens viseme patterns. This structural challenge causes CNN-based single-frame models to fall short even in limited word sets.

Despite the difficulty of viseme discrimination, it is not expected that only image-based CNN architectures will be used; a comparative study conducted by Tümer Sivri and her team showed that CNNs, which are successful in visual-spatial feature extraction, achieve lower performance than models such as BGRU that process temporal information. These results highlight how difficult it is to distinguish lip shapes phonologically in Turkish and reveal the inadequacy of purely image-based architectures in tasks requiring viseme discrimination. Therefore, the higher sentence recognition success observed in another study compared to word recognition success confirms that the solution to this structural viseme problem lies in shifting focus from individual lip shapes to temporal context and sequence information between words.

4.2 Existing Turkish Lip Reading Datasets

The biggest structural obstacle facing Turkish lip-reading research is the lack of extensive, standardized, and publicly available data sets compared to other languages. The

four studies analyzed have either created specialized and limited datasets or expanded existing small datasets to overcome this scarcity:

➢ **Creation of Specialized Datasets**

- All four studies have directly or indirectly contributed to the production of visual data specific to the Turkish language. These datasets were created using different word types and collection technologies:
- **Pourmousa and Özen:** Created a visual dataset of 71 words (adjectives, nouns, verbs) collected from 72 people.
- **Atila and Sabaz:** Created two new datasets with a broader scope, consisting of 111 words and 113 sentences.Yargıç and Doğan: Using an MS Kinect camera for the education of hearing-impaired children, they recorded 15 Turkish color names (e.g.,White, Brown, Red) five times from 10 individuals, creating a small dataset of 750 words.

➢ **Real-World Conditions and Data Augmentation Strategies**

To improve the generalization ability of small datasets, researchers have used methods that simulate challenging environmental conditions and increase the amount of data:

- **Tümer Sivri and others:** They used an existing dataset containing words/phrases from YouTube videos over 6 days and expanded the dataset from 1,390 to 5,560 examples using powerful image augmentation techniques such as sigmoidal transformation and horizontal flipping. This strategy aimed to include visual variation from different lighting and angles in the dataset.
- **Pourmousa and Özen:** They replicated their own small dataset by rotating the videos at different angles.
- **Yargıç and Doğan:** During the data collection process, they aimed to address 2D image problems (scale, angular change) caused by distance from the camera and head movements without the need for preprocessing, using the Kinect's 3D depth information.

Despite all these efforts, the datasets in all four analyzed studies are still limited, specialized, and small in size compared to the large-scale and standardized resources required for ALR systems to reach their full potential. This is a common and ongoing fundamental constraint in Turkish ALR research.

4.3 Related Works

The Turkish Automatic Lip Reading (ALR) literature reflects various methods and architectures developed to address the structural challenges of the language, despite limited data sources that lag behind global standards. Early approaches began with the work of

Alper Yargıç and Muzaffer Doğan (2013), representing the pre-deep learning era. This study laid the initial foundations for Turkish data collection by using a Microsoft Kinect camera and a traditional KNN classifier, based on the angles between 3D lip points. Subsequent studies focused on addressing the biggest gap in the field: data scarcity:

- **Hadi Pourmousa and Üstün Özen**: Created an original dataset of 71 words and increased the data using the video rotation (replication) method.
- **Hasan Atila and Ali Sabaz**: Created two more comprehensive datasets and demonstrated the effectiveness of temporal modeling by combining CNN and Bi-LSTM architectures.
- **Talya Tümer Sivri, Ali Berkol, and Hamit Erdem**: They proved that recurrent models such as BGRU and LSTM are more successful than CNNs that only extract visual-spatial features and applied powerful image augmentation techniques.

In summary, Turkish ALR literature continues to address the structural challenges of the language through strategies such as creating specialized datasets, data augmentation, and favoring temporal architectures. Early Turkish ALR studies, as well as those in other languages, have addressed simple tasks by combining visual feature extraction with classical machine learning methods. HMM (Hidden Markov Model) has been widely used in these studies [20, 21, 22], but it has been insufficient for long sequences and sentence-level recognition. Notable studies and models that have emerged with deep learning include:

- **Sarhan etc. (HLR-NET)**: A hybrid model consisting of Inception layers and BiGRU for word and character recognition [23].
- **Stafylakis & Tzimiropoulos**: 83% word accuracy using 3D CNN + LSTM on the BBC TV dataset [24].
- **Sterpu & Naomi:** ~54% success with DCT + AAM on the TCD-TIMIT dataset [25].
- **Thangthai etc.:** 48.9% accuracy on TCD-TIMIT with DNN + HMM [26].

Studies conducted at the sentence level have shown higher success with models that use temporal information:

- **Petridis etc.:** 91.8% sentence accuracy with RBM + Bi-LSTM on the AVIC and OuluVS2 datasets [27].
- **Thangthai etc.:** 84.7% success by adding lip-reading to speech recognition using the Kaldi DNN framework [28].
- **Huyen:** 88% accuracy using CNN + LSTM on a small German dataset [29].
- **Chen etc:** Sentence dataset containing 349 classes and 1705 characters for Mandarin, 61.2% accuracy with 3D CNN + DenseNet + Bi-LSTM [30].
- **Kurniawan & Suyanto:** 80% success with 3D CNN + BiGRU on an Indonesian dataset [31].

These studies demonstrate that deep learning models utilizing temporal information are more successful than classical methods, particularly at the sentence level, and form the basis for modern ALR systems.

5. Proposed Turkish Lip Reading System

The proposed system is designed to take into account the unique linguistic and articulation characteristics of Turkish. The architecture consists of four basic stages:

1. Preprocessing and Data Augmentation:

   Lip regions are detected and extracted from video frames. Data augmentation methods such as brightness adjustment, scaling, and temporal shifting are applied to increase robustness against speaker and environment variations.

2. Feature Extraction:

   3D-CNNs or hybrid CNN–Transformer structures extract spatial–temporal features associated with lip movements. This captures both short-term articulation patterns and longer-term movement structures .

3. Sequence Modeling:

   LSTM, GRU, or Transformer-based models convert visual features into sequences at the phoneme, syllable, or word level. These models are effective at learning long dependencies arising from the structure of Turkish.

4. Prediction and Language Model Integration:

   A Turkish language model is integrated into the system to improve prediction accuracy. The language model reduces vowel ambiguity by providing contextual constraints.

   The overall design is suitable for real-time inference and provides a solid foundation for assistive technologies, silent communication systems, and future multilingual VSR research.

6. Conclusion

This study provides an overview of visual speech recognition and addresses the fundamental challenges in developing a Turkish lip-reading system. While VSR research for English has advanced rapidly, Turkish remains understudied due to the complexity of its linguistic structure and the limited availability of datasets.

The proposed system combines modern spatio-temporal feature extraction methods adapted to Turkish with language model integration. This system aims to contribute to accessibility applications and provide a flexible, real-time framework for Turkish VSR.

Future work plans include increasing data set diversity, improving inference speed, and adding audiovisual models to the system to enhance robustness.

7. References

[1] M. A. Abrar et al., “Deep lip reading – A deep learning based lip-reading software for the hearing impaired,” 2019 IEEE R10 Humanitarian Technology Conference (R10-HTC), pp. 40–44, 2019. (link: https://doi.org/10.1109/R10-HTC47129.2019.9042439)

[2] M. Abishek et al., “Deep learning based lip reading for speech recognition,” 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT), pp. 1–6, 2024. (link: https://doi.org/10.1109/ICCCNT61001.2024.10725052)

[3] G. Tan et al., “Tackling event-based lip-reading by exploring multigrained spatiotemporal clues,” IEEE Transactions on Neural Networks and Learning Systems, vol. 36, no. 5, pp. 8279–8291, 2025. (link: https://doi.org/10.1109/TNNLS.2024.3440495)

[4] T. Afouras, J. S. Chung, and A. Zisserman, “ASR is all you need: Cross-modal distillation for lip reading,” ICASSP 2020 – IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 2143–2147, 2020. (link: https://doi.org/10.1109/ICASSP40776.2020.9054253)

[5] Z. Lin and N. Harte, “Uncovering the visual contribution in audio-visual speech recognition,” ICASSP 2025 – IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 1–5, 2025. (link: https://doi.org/10.1109/ICASSP49660.2025.10888423)

[6] P. Heracleous, H. Ishiguro, and N. Hagita, “Visual-speech to text conversion applicable to telephone communication for deaf individuals,” 2011 18th International Conference on Telecommunications, pp. 130–133, 2011. (link: https://doi.org/10.1109/CTS.2011.5898904)

[7] S. Petridis, J. Shen, D. Cetin, and M. Pantic, “Visual-only recognition of normal, whispered and silent speech,” 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6219–6223, 2018. (link: https://doi.org/10.1109/ICASSP.2018.8461596)

[8] H. Kode, B. S. S. Kotipalli, H. R. Cheruku, and S. S. Gadde, “A survey on deep learning-based approaches for automated lip reading,” 2025 IEEE 2nd International Conference on Deep Learning and Computer Vision (DLCV), pp. 1–6, 2025. (link: https://doi.org/10.1109/DLCV65218.2025.11088852)

[9] P. Sindhura, S. J. Preethi, and K. B. Niranjana, “Convolutional neural networks for predicting words: A lip-reading system,” 2018 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT), pp. 929–933, 2018. (link: https://doi.org/10.1109/ICEECCOT43722.2018.9001505)

[10]Hunter, H. S. (2013). *A STUDY OF TURKISH VOWEL HARMONY AND THE POWER OF LANGUAGE* [Undergraduate thesis, Kent State University]. OhioLINK Electronic Theses and Dissertations Center. http://rave.ohiolink.edu/etdc/view?acc_num=ksuhonors1376315922

[11]Berkol A, Tümer-Sivri T, Pervan-Akman N, Çolak M, Erdem H. Visual Lip Reading Dataset in Turkish. Data. 2023; 8(1):15. https://doi.org/10.3390/data8010015

[12]N. Radha, A. Shahina and A. N. khan, "A Survey on Visual Speech Recognition Approaches," 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), Coimbatore, India, 2021, pp. 934-939, doi: 10.1109/ICAIS50930.2021.9395878. keywords: {Visualization;Lips;Working environment noise;Speech recognition;Feature extraction;History;Visual databases;Region of Interest;Hidden Markov Model;Active Shape Model;Motion History Image;Audio-Visual},

[13]T. Saitoh and R. Konishi, "Profile Lip Reading for Vowel and Word Recognition," 2010 20th International Conference on Pattern Recognition, Istanbul, Turkey, 2010, pp. 1356-1359, doi: 10.1109/ICPR.2010.335. keywords: {Feature extraction;Nose;Euclidean distance;Face;Mouth;Hidden Markov models;Lips;lip reading;profile;vowel recognition;word recognition},

[14] IBM, "What is Transfer Learning?," IBM Think. [Online]. Available: https://www.ibm.com/think/topics/transfer-learning. [Accessed: Nov. 28, 2025].

[15] T. Stafylakis ve G. Tzimiropoulos, “Combining Residual Networks with LSTMs for Lipreading,” arXiv e-prints, 2017, Art. no. arXiv:1703.04105.

[16] C. Sheng et al., “Deep Learning for Visual Speech Analysis: A Survey,” arXiv e-prints, 2024, Art. no. arXiv:2205.10839v2.

[17] Pourmousa, H.; Özen, Ü. Lip reading using deep learning in Turkish language. *IAES Int. J. Artif. Intell.* 2024, *13*, 3250–3261.
[18] Atila, Ü.; Sabaz, F. Turkish lip-reading using Bi-LSTM and deep learning models. *Eng. Sci. Technol. Int. J.* **2022**, *35*, 101206.
[19] Tümer Sıvrı, T.; Berkol, A.; Erdem, H. Lip Reading Using Various Deep Learning Models with Visual Turkish Data. *GUJ Sci* 2024, *37*, 1190–1203.

[20]S.S. Morade, S. Patnaik

A novel lip reading algorithm by using localized ACM and HMM: Tested for digit recognition

Optik (Stuttg), 125 (2014), pp. 5181-5186, 10.1016/J.IJLEO.2014.05.011

[21]A. Yargic, M. Dogan, A lip reading application on MS Kinect camera, in: 2013 IEEE Int Symp Innov Intell Syst Appl IEEE INISTA 2013, 2013. https://doi.org/10.1109/INISTA.2013.6577656.

[22]N. Puviarasan, S. Palanivel Lip reading of hearing impaired persons using HMM

Expert Syst. Appl., 38 (4) (2011), pp. 4477-4481

[23]A.M. Sarhan, N.M. Elshennawy, D.M. Ibrahim
HLR-Net: A hybrid lip-reading model based on deep convolutional neural networks
Comput. Mater. Contin., 68 (2021), pp. 1531-1549, 10.32604/CMC.2021.016509

[24] T. Stafylakis, G. Tzimiropoulos, Combining Residual Networks with LSTMs for Lipreading, in: Proc Annu Conf Int Speech Commun Assoc INTERSPEECH 2017-August, 2017, 3652–3656. Google Scholar

[25]G. Sterpu, H. Naomi, Towards Lipreading Sentences with Active Appearance Models. AVSP, 2017, 70–75.Google Scholar
[26]K. Thangthai, R. Harvey, Improving computer lipreading via DNN sequence discriminative training techniques, in: Proc Annu Conf Int Speech Commun Assoc INTERSPEECH 2017-August, 2017, pp. 3657–3661. https://doi.org/10.21437/INTERSPEECH.2017-106.
[27]S. Petridis, Y. Wang, Z. Li, M. Pantic End-to-End Audiovisual Fusion with LSTMs The 14th International Conference on Auditory-Visual Speech Processing. International Speech Communication Association (2018), pp. 36-40 Google Scholar

[28] K. Thangthai, R.W. Harvey, S.J. Cox, B.J. Theobald, Improving lip-reading performance for robust audiovisual speech recognition using DNNs, in: AVSP, 2015, pp. 127–131.Google Scholar
[29] C.T. Huyen German Word Level Lip Reading with Deep Learning Hochschule für angewandte Wissenschaften Hamburg (2019) Doctoral dissertation Google Scholar
[30] X. Chen, J. Du, H. Zhang Zhang H (2020) Lipreading with DenseNet and resBi-LSTM Signal Image Video Process, 14 (5) (2020), pp. 981-989 CrossrefView in ScopusGoogle Scholar

[31] A. Kurniawan, S. Suyanto, Syllable-Based Indonesian Lip Reading Model, in: 2020 8th Int Conf Inf Commun Technol ICoICT 2020, 2020. https://doi.org/10.1109/ICOICT49345.2020.9166217.

Other References

S. Debnath, P. Roy, V. Justin, S. Naik, "Study of different feature extraction method for visual speech recognition," 2021 International Conference on Computer Communication and Informatics (ICCCI). [Online]. Available: https://www.researchgate.net/publication/351645818_Study_of_different_feature_extraction_method_for_visual_speech_recognition. [Accessed: Nov. 26, 2025].

M. K. Mistry, "Comprehensive feature extraction for the recognition of visual speech and speakers," International Research Journal of Natural and Applied Sciences (AARF), Vol. 4, Issue 1, Jan. 2017. [Online]. Available: https://www.aarf.asia/current/2025/Feb/sFuIDX7jH7eNjCV.pdf. [Accessed: Nov. 26, 2025].

S. Das, A. Tariq, T. Santos, S. S. Kantareddy, I. Banerjee, "Recurrent Neural Networks (RNNs): Architectures, Training Tricks, and Introduction to Influential Research," in *Machine Learning for Brain Disorders*, Neuromethods, vol. 197, Humana (SpringerLink), pp. 117–138, First Online: 23 July 2023. [Online]. Available: https://link.springer.com/protocol/10.1007/978-1-0716-3195-9_4. [Accessed: Nov. 27, 2025].

Y. M. Assael, B. Shillingford, S. Whiteson, N. de Freitas, "LipNet: End-to-End Sentence-level Lipreading," arXiv:1611.01599, submitted Nov. 5, 2016; revised Dec. 16, 2016. [Online]. Available: https://arxiv.org/abs/1611.01599. [Accessed: Nov. 27, 2025].

M. Hao et al., “A Survey of Research on Lipreading Technology,” IEEE Access, vol. 8, pp. 204518–204544, 2020. doi: 10.1109/ACCESS.2020.3036865.

Figures

[1] Ipsic, I. (ed.), *Speech and Language Technologies*. BoD – Books on Demand / Google Books. [Online]. Available:

https://books.google.com.tr/books?hl=en&lr=&id=nHSfDwAAQBAJ&oi=fnd&pg=PA279&dq=Visual+Speech+Recognition+Techniques&ots=hM3X33Klpk&sig=avVQSv8HiM1xEGTQeGiredvQFlo&redir_esc=y#v=onepage&q=Visual%20Speech%20Recognition%20Techniques&f=false.

[Accessed: Nov. 28, 2025].