A NOVEL TASK-ORIENTED APPROACH TOWARD AUTOMATED LIP-READING SYSTEM IMPLEMENTATION

D. Ivanko, D. Ryumin

St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences, St. Petersburg
Federal Research Center of the Russian Academy of Sciences, SPC RAS, Saint-Petersburg, Russian Federation – denis.ivanko@uni-ulm.de, ryumin.d@iias.spb.su

Commission II, WG II/5

KEY WORDS: Automated lip-reading, Deep neural networks, Hidden Markov models, Geometric features, Region-of-interest detection.

ABSTRACT:

Visual information plays a key role in automatic speech recognition (ASR) when audio is corrupted by background noise, or even inaccessible. Speech recognition using visual information is called lip-reading. The initial idea of visual speech recognition comes from humans’ experience: we are able to recognize spoken words from the observation of a speaker's face without or with limited access to the sound part of the voice. Based on the conducted experimental evaluations as well as on analysis of the research field we propose a novel task-oriented approach towards practical lip-reading system implementation. Its main purpose is to be some kind of a roadmap for researchers who need to build a reliable visual speech recognition system for their task. In a rough approximation, we can divide the task of lip-reading into two parts, depending on the complexity of the problem. First, if we need to recognize isolated words, numbers or small phrases (e.g. Telephone numbers with a strict grammar or keywords). Or second, if we need to recognize continuous speech (phrases or sentences). All these stages disclosed in detail in this paper. Based on the proposed approach we implemented from scratch automatic visual speech recognition systems of three different architectures: GMM-CHMM, DNN-HMM and purely End-to-end. A description of the methodology, tools, step-by-step development and all necessary parameters are disclosed in detail in current paper. It is worth noting that for the Russian speech recognition, such systems were created for the first time.

1. INTRODUCTION

The initial idea of visual speech recognition comes from humans’ experience: we are able to recognize spoken words from the observation of a speaker's face without or with limited access to the sound part of the voice. Visual information plays a key role in automatic speech recognition (ASR) when audio is corrupted by background noise, or even inaccessible. Speech recognition using visual information is called lip-reading.

However, the history of automatic visual speech recognition began only decades ago. During the last two decades, there have been significant advances in the research of audio-based ASR. Initially, researchers expected that automatic visual speech recognition would be easily accomplished based on the progress achieved in development of audio-based ASR. However, early attempts did not yield good results. Despite the poor performance, visual features still helped boost the ASR performance on some low-quality audio data through audio-visual speech information fusion. The active development of the field of automatic speech recognition in the modern world has brought human-machine interfaces to a new level.

The use of a natural form of speech communication with machines greatly facilitates implementation of a huge number of different tasks for people, e.g. from casual (smart home speech recognition, dialling a phone, sending a message by speech command) to business and industry (automatic call processing in call centres, information kiosks at airports, lobbies of railway stations, etc.). Today, in quiet office environments, for a variety of tasks speech recognition can approach almost hundred percent of accuracy. However, it is often achieved under the condition of a limited vocabulary and a stricted grammar. Nonetheless, one of the most difficult tasks in the field of automatic speech recognition is the recognition of continuous speech. It should be noted that the existing systems and models of automatic recognition of continuous speech are still significantly inferior to the speech abilities of a human, especially in real conditions of use. The main difficulty of this problem lies in the great variability of the basic parameters of speech, which are influenced by many factors. First of all, it is a random component of the speech formation process. Due to the individuality of the speech-forming apparatus of people, the same statement uttered by different speakers may differ significantly both in the time and in the frequency ranges. Gender, age, accent, emotional and physical state of the speaker have a significant impact and complicate the task of effective speech recognition for automated systems. In addition, the acoustic effect has a great influence on the speech recognition accuracy - the type of microphone, its characteristics, location relative to the speaker's mouth, the surrounding acoustic environment (presence of external noise, reverberation, etc.). Thus, in many operating conditions, the existing automatic recognition systems cannot provide the required quality of recognition even when using various methods of noise reduction and filtering (Almajai et al., 2016).

At the same time, we must not forget about the way that people themselves use to better understand the interlocutor's statements in difficult acoustic conditions or when several people talk at the same time. Oral speech is transmitted both through the acoustic channel (through sound) and through the visual channel (lip movements). Back in 1976, it was proved that natural speech is generated multimodally and transmitted simultaneously over several channels in the form of audio and video information, which duplicates and complements each other, helping to correctly perceive speech in many difficult situations. A similar phenomenon was called the McGurk-MacDonald effect and marked the beginning of the development of a new field - the automatic lip-reading. Since the early 90s, there have been several attempts to use visual information about speech in addition to acoustic information, to improve the accuracy and reliability of automatic recognition systems. In a number of studies, the developed audio-visual speech recognition systems have demonstrated better

recognition results than uni-modal systems. However, compared to the active development of the field of acoustic speech recognition, there is not much research on visual speech recognition to date (Ivanko et al., 2016). Despite the great potential of the use of visual information about speech, this area still has a large number of unresolved problems.

Today, there is no generally accepted approach to the development of visual speech recognition systems (Wang et al., 2019). There are no publicly available representative databases for training models that would have all the necessary parameters, such as a sufficient number of speakers, phonemic-viseme temporal labelling, vocabulary size adequate to the task (Morade et al., 2015), etc. (there are practically no public databases for languages other than English). There is no research into the effect of video recording speed on speech recognition accuracy (Thangthai et al., 2015). There is little research on the effect of acoustically noisy environments on the performance of visual speech recognition systems, and quite a few studies have focused on inflectional languages (such as Russian). However, there is a huge difference between the recognition of analytical languages (for example, English) and inflected languages, due to the presence in the latter of a much larger number of word forms and grammatical rules.

2. BACKGROUNDS AND RELATED RESEARCH

Nowadays, automatic speech recognition systems based on the processing of video signals show impressive results (Fernandez-Lopez, A. et al., 2018). In general, researchers divide the solution of the visual speech recognition problem into two parts. First, in extracting the most informative features from video modality and second, in finding the most successful way to train the recognition model (Zhou et al., 2014).

In a rough approximation, we can divide the task of lip-reading into two parts, depending on the complexity of the problem (Kashevnik et al., 2021). First, if we need to recognize isolated words, numbers or small phrases (e.g. Telephone numbers with a strict grammar or keywords). Or second, if we need to recognize continuous speech (phrases or sentences). Based on this information we have to choose different size of a database, different methods for features extraction and model training (Ryumin et al., 2019).

Automatic visual speech recognition as a research field lies at the intersection of several areas of knowledge: digital signal processing, computer vision, statistical modeling, machine learning, etc. Thus, the development of visual speech recognition (VSR) system involves not only the correct implementation of all components related to those fields, but also their proper integration, which is sometimes a non-trivial task (Kagirov et al., 2020). In the general case, the methods used in this study can be divided into two subgroups: (1) methods of signal preprocessing and informative features extraction and (2) methods of statistical modeling and machine learning.

To date, the are several widely used approaches toward visual speech recognition task (lip-reading). As well as commonly used visual features extraction algorithms and visual speech modeling methods. In addition, we started with investigation of region-of-interest (ROI) detection approaches (Ivanko et al., 2018a). We found out that Active Appearance Models-based and Haar-like features-based methods most widely used for this purpose. As for visual features extraction methods, we have discovered that to date there is no universally accepted feature set for representing visual speech information. We briefly considered the main challenges in the area (Katsaggelos et al., 2015), such as speaker dependency, pose variation and temporal information. We identified the most prospective approaches for visual features extraction, namely pixel-based and geometry-based methods (Lin et al., 2017). As round-up, we pointed out the most widely used visual speech modeling methods, which turned out to be support vector machines (SVM), hidden Markov models (HMM) or deep neural networks (DNN) based approaches (Lu et al., 2020). In this paper we present the developed task-oriented approach for creating practical visual speech recognition systems. It is based on conducted experiments and analysis of the field of research with its main goal to be some kind of a roadmap for researchers who need to build a reliable and robust audio-visual speech recognition system. We highlighted that the problem of automatic visual speech recognition lies, firstly, in extracting the most informative features from video modality and, secondly, in the most successful way of training the recognition models. We also noted the fact, that despite significant successes recently achieved in the field of acoustic speech recognition, the task of automated lip-reading still remains underdeveloped.

3. DATA COLLECTION AND ANALYSIS

According to our analysis, to date, there are 47 different publicly available visual or audio-visual speech databases that are found in the scientific literature at the time of research (Fernandez-Lopez et al., 2017). Considered databases were compared by the following parameters: language, year of collection, number of subjects, total number of samples, video resolution and fps, audio sampling rate, availability for researchers.

Based on the conducted analysis, the following conclusions can be made:

- The vast majority of existing datasets are designed for the English language (35 out of 47). And no more than 1-2 datasets for other languages, including German, French, Czech, Japanese, Chinese, Spanish, Polish and Russian.

- The maximum frame rate of video is 100 fps in MODALITY and NDUTAVSC datasets. In RM-3000 dataset the recording speed is 60 fps and in three others (AVLetters2, AGH AV, VLRF) it is 50 fps. In all the rest (42), the video recording speed does not exceed 25-30 frames per second.

- Only about half of the reviewed databases (22) are designed to work with continuous speech recognition tasks (sentences). The rest were collected to solve simpler tasks, such as recognition of isolated numbers, letters or words.

- The vast majority of existing databases contain records from 10 to 50 speakers (Lu et al., 2019). However, there are also databases in which the number of speakers is clearly not enough to build speaker-independent recognition systems: RM-3000 (1 speaker), QuLips (2 speakers), AVLetters2 (5 speakers). etc.

- Not all databases have suitable video resolution. E.g., LRS (160 × 160), LSVSR (128 × 128), LRW (256 × 256) datasets contains only recordings of speakers' mouth, ignoring the rest of the face. Datasets such as AVLetters (376 × 288), IBMSR (368 × 240), VIDTIMIT (512 × 384) simply have a fairly low resolution of video.

Thus, according to our findings, we choose to test our approach on HAVRUS (Verkhodanova et al., 2016) dataset for small phrases recognition task and GRID (Cookie at al., 2008) dataset for sentence level recognition task. The description of the main characteristics of abovementioned corpora are presented in the Figure 1.

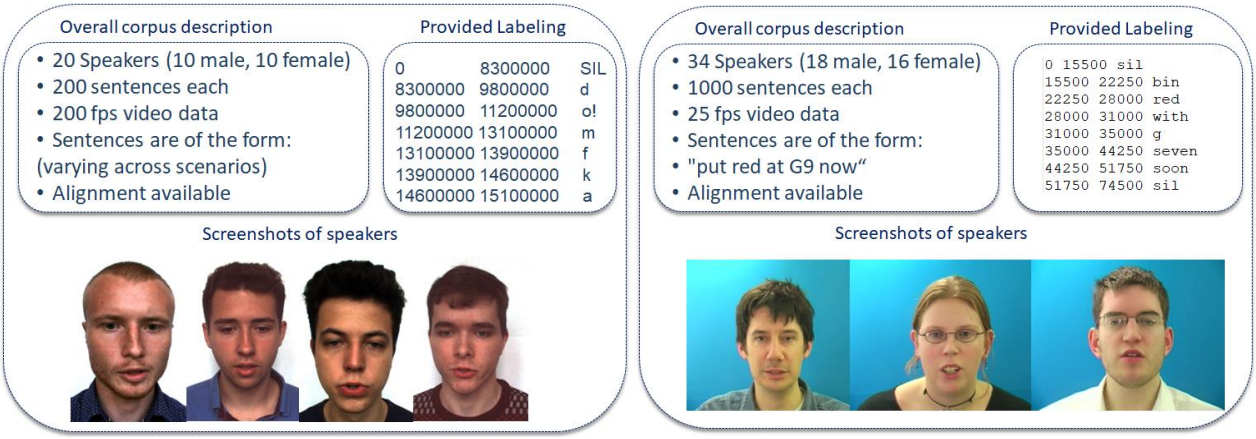

Figure 1 – HAVRUS (left) and GRID (right) visual speech datasets characteristics

4. PROPOSED TASK-ORIENTED APPROACH

Based on the conducted experimental evaluations as well as on analysis of the research field we propose a novel task-oriented approach towards practical lip-reading system implementation. In a rough approximation, we can divide the task of lip-reading into two parts, depending on the complexity of the problem. First, if we need to recognize isolated words, numbers or small phrases (e.g. telephone numbers with a strict grammar or keywords). Second, if we need to recognize continuous speech (phrases or sentences). Based on this information we have to choose different size of a database, different methods for features extraction and model training (Lu et al., 2018). All these main stages depicted in the figure 1. E.g., if our goal is to build automated spelling system, we can choose small or medium size AVLetters (Matthews et al., 2002) or AVLetter2 (Cox et al., 2008) dataset (or collect similar in a target language), use AAM for ROI detection and geometry-based features extraction, followed by HMM for model training, as was done, for example, in the work (Bear et al., 2017). If our task is to build an isolated digit phone dialing system, we have to choose XM2VTS (Messer et al., 1999), CUAVE (Patterson et al., 2002) or similar database, then use Haar-based ROI detection techniques and DCT-LDA features extraction methods, followed by either SVM or MS-HMM for model training, as was done in the works (Stewart et al., 2016). If, instead of building phone dialing system, we want to create reliable phone number recognition system of multiple speakers, we have to choose medium size dataset (e.g., such as HAVRUS, as was done in current research). Use Haar cascades or AAM for ROI detection, and extract combination of pixel and geometry-based features, followed by training of CHMM models (Ivanko et al., 2018).

According to the proposed approach (Figure 1), the general steps in creating practical lip-reading application will be:

(1) depending on the task, choose or collect appropriate dataset for training. Small or medium size for isolated digits, letters, words or restricted grammar phrases recognition or large size databases for sentence level recognition.

(2) both methods of ROI detection (Haar feature based or AAM based) are commonly used in automated lip-reading task (Howell et al., 2016). Our experiments also did not show superiority of any of this method over another. Thus, the both methods are equally applicable.

(3) according to our results regarding informative features extraction the combination of both pixel-based and geometry-based features provides the best recognition results on the first type of tasks (Ivanko et al., 2019). However, on the sentence level recognition task with sufficient amount of training data, NN-based methods (especially CNN-based) show better results (Lee et al., 2016). Some researchers also combined such NN features with traditional ones (Chung et al., 2016), but this usually was used to boost performance when struggling to properly train a network due to the lack of data

(4) for the model training, GMM-HMMs and their modifications (Coupled hidden Markov models, CHMM or multistream hidden Markov models, MSHMM (Ivanko et al., 2018b)) are the first choice when building small vocabulary lip-reading systems. SVM also applicable, if the dictionary not exceed several units (Ivanko et al., 2020). However, for the continuous visual speech recognition, the HMM-DNN methods or HMM-RNN methods (such as long-short term memory, LSTM or gated recurrent unit, GRU) demonstrated much better recognition results.

In current series of experiments, we used two different datasets for model training. GRID dataset for English language and HAVRUS dataset for Russian language. Using these datasets, we trained AVSR's of tree different topologies, namely GMM-CHMM, DNN-HMM and End-to-end.

	GMM-CHMM	DNN-HMM	End-to-end
GRID	55.12	71.34	84.3
HAVRUS	45.18	25.57	1.13

Table 1. Word Recognition Rate (WRR, %) comparison of different implemented VSR systems.

As we can observe from obtained results (Table 1) on the video part of the GRID corpus, end-to-end approach (84.3%) outperforms GMM-CHMM (55.12%) and DNN-HMM (71.34%) ones. This result is quite expected and once again confirms the superiority of neural network approaches compared to the others in conditions when we have enough data to effectively train NN models. However, the situation drastically changes when we train same models on the video part of the Russian HAVRUS corpus. In this case, the best recognition results were obtained using traditional GMM-CHMM approach (45.18%), and the results on NN topologies were much worse. These recognition results can be explained be the fact that HAVRUS corpus is simply much smaller (4000

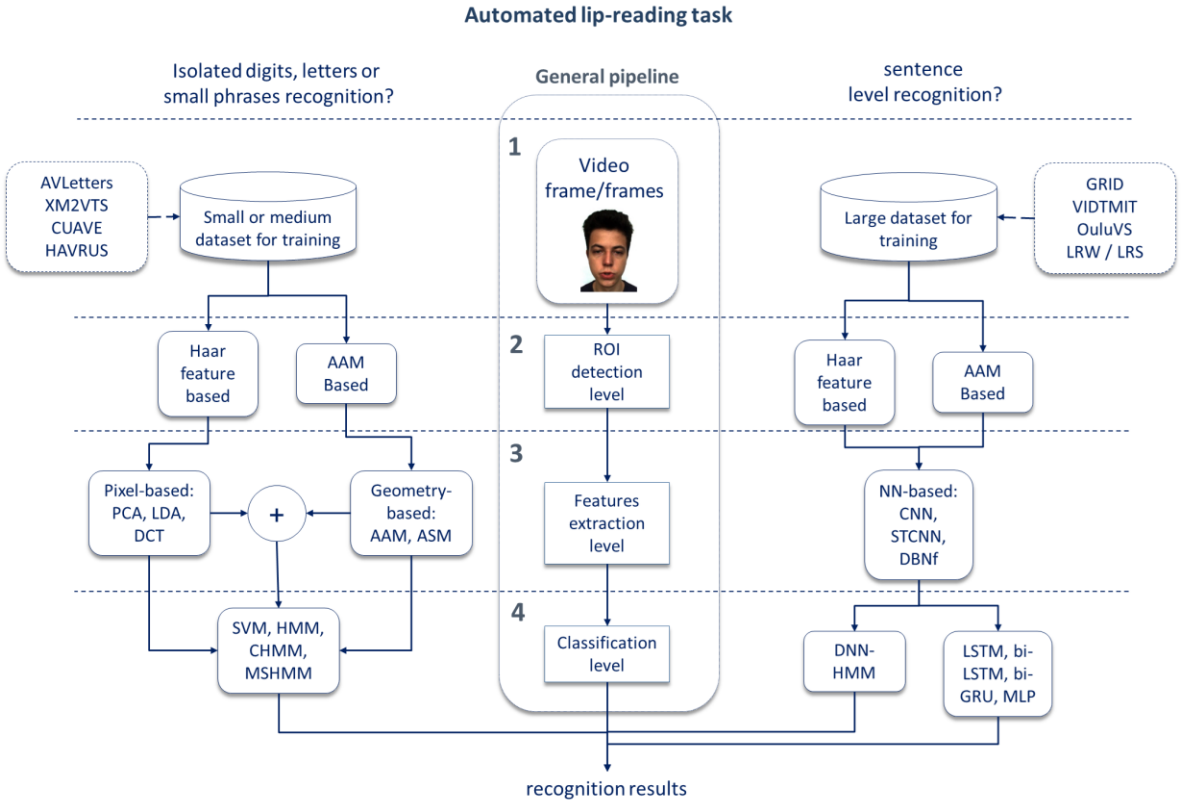

Figure 2 – Proposed task-oriented approach toward practical lip-reading system implementation

phrases in total) than the GRID one (34000 phrases in total) and these data amount is not sufficient for effective training of NN models.

5. CONCLUSIONS

In this paper we present the developed task-oriented approach for creating practical visual speech recognition systems. It is based on conducted experiments and analysis of the field of research with its main goal to be some kind of a roadmap for researchers who need to build a reliable and robust audio-visual speech recognition system.

We highlighted that the problem of automatic audio-visual speech recognition lies, firstly, in extracting the most informative features from each modality and, secondly, in the most successful way of fusion both modalities. We also noted the fact, that despite significant successes recently achieved in the field of acoustic speech recognition, the task of automated lip-reading still remains underdeveloped.

By this work we tried to lay the foundation, outline certain boundaries and highlight some milestones of the field. However, there are still a lot of questions regarding visual speech recognition that need to be answered. In the following we propose several future directions for further hypothetical improvements of automatic audio-visual Russian speech recognition.

- the influence of visual noise was left out of the scope of our work. Nevertheless, this is of great practical interest, since, in contrast to quiet office conditions, in real-world applications, the variability of lighting plays a key role in the task of automated lip-reading.

- Extending the existing and recording new visual speech databases is also of great practical interest. Because of that many experiments were limited to traditional GMM-HMM models. Such a problem can be resolved by increasing the amount of data available for training.

- An extensive usage of deep learning approach alongside with the collection of additional data might further improve the performance of recognition models.

ACKNOWLEDGEMENTS

This research is financially supported by the Russian Foundation for Basic Research (project No. 19-29-09081 мк).

REFERENCES

Almajai, I., Cox, S., Harvey, R., & Lan, Y. (2016). Improved speaker independent lip reading using speaker adaptive training and deep neural networks. In 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2722-2726.

Bear, H.L., Harvey, R. (2017). Phoneme-to-viseme mappings: the good, the bad, and the ugly, Speech Comm. 95, pp. 40–67.

Chung J. S., Zisserman, A. (2017). Lip reading in the wild. *In Proceedings of ACCV*, pp. 87-103.

Cooke, M., Barker, J., Cunningham, S., Shao X. (2008). An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, 120 (5), pp. 2421–2424.

Cox, S., Harvey, R., Lan, Y., Newman, J., Theobald, B. (2008) The challenge of multispeaker lip-reading. In Proceedings of

International conference Auditory-Visual Speech Processing (AVSP), pp. 179-184.

Fernandez-Lopez, A., & Sukno, F. M. (2018). Survey on automatic lip-reading in the era of deep learning. Image and Vision Computing, 78, 53-72.

Fernandez-Lopez, A., Martinez, O., & Sukno, F. M. (2017). Towards estimating the upper bound of visual-speech recognition: The visual lip-reading feasibility database. In 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), pp. 208-215.

Howell, D., Cox, S., & Theobald, B. (2016). Visual units and confusion modelling for automatic lip-reading. Image and Vision Computing, 51, 1-12.

Ivanko, D., Karpov, A. (2016). An analysis of perspectives for using high-speed cameras in processing dynamic video information. In: Journal of SPIIRAS proceedings, 44(1), pp. 98-113.

Ivanko, D., Karpov, A., Fedotov, D., Kipyatkova, I., Ryumin, D., Ivanko, Dm., Minker, W., Zelezny, M. (2018a) Multimodal speech recognition: increasing accuracy using high-speed video data. *Journal of Multimodal User Interfaces*, 12(4), pp. 319-328.

Ivanko, D., Ryumin, D., Axyonov, A., Zelezny, M. (2018b). Designing advanced geometric features for automatic Russian visual speech recognition. In: *Proceedings of 20th International Conference on Speech and Computer, SPECOM 2018*, Leipzig, Germany, September 18-22, LNAI 11096, pp. 245-255.

Ivanko, D., Ryumin, D., Kipyatkova, I., Axyonov, A., Karpov, A. (2019). Lip-reading using pixel-based and geometry-based features for multimodal human-robot interfaces. In: 14th International Conference on Electromechanics and Robotics “Zavalishin’s Readings”, ERZR 2019. *Proc. of Smart Innovation, Systems and Technologies book series of Springer*, pp. 477-486.

Ivanko, D., Ryumin, D., Karpov, A. (2020). An Experimental Analysis of Different Approaches to Audio-Visual Speech Recognition and Lip-Reading. In: 15th International Conference on Electromechanics and Robotics “Zavalishin’s Readings”, ERZR 2020. *Proc. of Smart Innovation, Systems and Technologies book series of Springer*, pp. 197-209.

Kagirov, I., Ivanko, D., Ryumin, D., Axyonov, A., & Karpov, A. (2020). TheRuSLan: Database of Russian Sign Language. In Proceedings of The 12th Language Resources and Evaluation Conference, pp. 6079-6085.

Kashevnik, A. et al., "Multimodal Corpus Design for Audio-Visual Speech Recognition in Vehicle Cabin," in IEEE Access, doi: 10.1109/ACCESS.2021.3062752.

Katsaggelos, K., Bahaadini, S., Molina, R. (2015). Audiovisual Fusion: Challenges and New Approaches. *In Proceedings of the IEEE*, 103 (9), 1635-1653.

Lee, D., Lee, J., & Kim, K. E. (2016). Multi-view automatic lip-reading using neural network. In Asian conference on computer vision, pp. 290-302.

Lin, B. S., Yao, Y. H., Liu, C. F., Lien, C. F., & Lin, B. S. (2017). Development of novel lip-reading recognition algorithm. IEEE Access, 5, 794-801.

Lu, Y., Yan, J., & Gu, K. (2018). Review on automatic lip-reading techniques. International Journal of Pattern Recognition and Artificial Intelligence, 32(07), 1856007.

Lu, Y., & Li, H. (2019). Automatic lip-reading system based on deep convolutional neural network and attention-based long short-term memory. Applied Sciences, 9(8), 1599.

Lu, Y., & Yan, J. (2020). Automatic lip reading using convolution neural network and bidirectional long short-term memory. International Journal of Pattern Recognition and Artificial Intelligence, 34(01), 2054003.

Matthews, I., Cootes, T.F., Bangham, J.A., Cox, S., Harvey, R. (2002). Extraction of visual features for lipreading, IEEE Trans. Pattern Anal. Mach. Intell. 24 (2), pp. 198–213.

Messer, K., Matas, J., Kittler, J., Luettin, J., Maitre, G. (1999). XM2VTSDB: the extended M2VTS database, Proc. International Conference on Audio and Video-based Biometric Person Authentication, vol. 964, pp. 965–966.

Morade, S. S., & Patnaik, S. (2015). Comparison of classifiers for lip reading with CUAVE and TULIPS database. Optik, 126(24), pp. 5753-5761.

Patterson, E., Gurbuz, S., Tufekci, Z., Gowdy, J. (2002). CUAVE: a new audio-visual database for multimodal human-computer interface research. In Proceedings of IEEE ICASSP 2002, pp. 2017-2020.

Ryumin, D., Ivanko, D., Axyonov, A., Kagirov. I., Karpov. A., Zelezny, M. (2019). Human-robot interaction with smart shopping trolley using sign language: data collection. In: 2019 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops), Kyoto, 11-15 march, pp. 949-954.

Stewart, D., Seymour, R., Pass, A., Ming, J. (2016). Robust audio-visual speech recognition under noisy audio-video conditions. In IEEE Transactions on Cybernetics, 44(2), pp. 175–184.

Thangthai, K., Harvey, R. W., Cox, S. J., & Theobald, B. J. (2015, September). Improving lip-reading performance for robust audiovisual speech recognition using DNNs. In AVSP, pp. 127-131.

Verkhodanova, V., Ronzhin, A., Kipyatkova, I., Ivanko, D., Karpov, A., Zelezny, M. (2016). HAVRUS corpus: high-speed recordings of audio-visual Russian speech. In: *Proceedings of 18th International Conference on Speech and Computer, SPECOM 2016*, August 23-27, Budapest, Hungary, LNAI 9811, pp. 338-345.

Wang, C. (2019). Multi-grained spatio-temporal modeling for lip-reading. arXiv preprint arXiv:1908.11618.

Zhou, Z., Hong, X., Zhao, G., Pietikainen M. (2014) A compact representation of visual speech data using latent variables. *IEEE Transactions on Pattern Analysis and Machine Intelligent*, 36(1), pp. 181–187.