Image Anal Stereol 2018;36:159-171
Original Research Paper

doi: 10.5566/ias.1859

LIP-READING VIA DEEP NEURAL NETWORKS USING HYBRID VISUAL FEATURES

FATEMEH VAKHSHITEH[1], FARSHAD ALMASGANJ✉,[1], AHMAD NICKABADI[2]

[1]Department of Biomedical Engineering, Amirkabir University of Technology; [2]Department of Computer Engineering and IT, Amirkabir University of Technology, Iran
e-mail: f.vakhshiteh@aut.ac.ir, almas@aut.ac.ir, nickabadi@aut.ac.ir

(Received November 18, 2017; revised April 29, 2018; accepted May 8, 2018)

ABSTRACT

Lip-reading is typically known as visually interpreting the speaker's lip movements during speaking. Experiments over many years have revealed that speech intelligibility increases if visual facial information becomes available. This effect becomes more apparent in noisy environments. Taking steps toward automating this process, some challenges will be raised such as coarticulation phenomenon, visual units' type, features diversity and their inter-speaker dependency. While efforts have been made to overcome these challenges, presentation of a flawless lip-reading system is still under the investigations. This paper searches for a lip-reading model with an efficiently developed incorporation and arrangement of processing blocks to extract highly discriminative visual features. Here, application of a properly structured Deep Belief Network (DBN)-based recognizer is highlighted. Multi-speaker (MS) and speaker-independent (SI) tasks are performed over CUAVE database, and phone recognition rates (PRRs) of 77.65% and 73.40% are achieved, respectively. The best word recognition rates (WRRs) achieved in the tasks of MS and SI are 80.25% and 76.91%, respectively. Resulted accuracies demonstrate that the proposed method outperforms the conventional Hidden Markov Model (HMM) and competes well with the state-of-the-art visual speech recognition works.

Keywords: Deep belief Networks, Hidden Markov Model, lip-reading, Restricted Boltzmann Machine

INTRODUCTION

Lip-reading is typically known as visually interpreting the movements of speaker's lip during speaking. From mathematical point of view, it requires converting the mouth movements to a truthful representation for possible visual recognition (Potamianos *et al.*, 2003). Experiments over many years have revealed that speech intelligibility increases if both the audio and visual information are available (McClain *et al.*, 2004, Sumby and Pollack, 1954). This effect becomes more apparent in noisy conditions like hospital emergency rooms. In some works, the visual information extracted from lip movements are integrated into Automatic Speech Recognition (ASR) systems to contribute to its robustness and accuracy; this is especially evaluated when the audio stream is corrupted by different levels of noises (Potamianos *et al.*, 2003).

Taking a step toward automating the lip-reading process, some challenges will appear compared to the conventional audio speech recognition tasks. Firstly, the coarticulation phenomenon can cause the visible speech articulators to be in different positions for the same underlying sound, and thus, the visual features become more dependent to the context (Lan *et al.*, 2012a). In some studies, viseme (facial and lip position corresponding to a particular phoneme) classes are used to resolve this issue (Bowden *et al.*, 2012, Savchenko and Khokhlova, 2014). However, other issues such as the homophone effect, insufficient training data per class, and substantial lack of distinction between classes, when there are too many visemes within a set, are faced in real conditions. Secondly, the choice of features for visual speech recognition purposes is rather wide. In the literature, several types of visual features are evaluated, which have been commonly categorized as those depending on pixels and those based on the models (Matthews *et al.*, 2002). Although some efforts have been made in this regard, extraction of optimized visual features is still under the investigation (Lan *et al.*, 2009, Lan *et al.*, 2010). Thirdly, selected visual features are data-driven and so are greatly speaker-dependent. Hence, many of the presented automated lip-reading systems are either using speaker-dependent or multi-speaker configuration.

Finding a method to face all these challenges in unison is still under the investigations. In the current study, it is aimed to introduce an approach which performs the major parts of a lip-reading system, but the main emphasis is over the feature extraction part. This is done by developing an efficient incorporation and arrangement of certain function blocks, by which highly informative yet discriminative visual features are extracted from lip images.

Proposed feature extraction pipeline is arranged in a way that pass the speakers' frame sequences through the function blocks of face detection, lip localization, scaling and reshaping, and meaningful feature extraction by a combination of Principal Component Analysis (PCA), Active Shape Model (ASM), and geometric feature calculations. The feature extraction pipeline is further improved by normalization, addition of Linear Discriminant Analysis (LDA) technique, Maximum Likelihood Linear Transform (MLLT), and Speaker Adaptive Training (SAT), and augmentation of dynamic information. The novelty of this research is the arrangement and newly inclusion of some function blocks in way that resulted in appropriate "hybrid" visual features. The advantage of the proposed feature extraction over some state-of-the art methods is that it is not facing the challenges of robustly tracking the lip-contour and hand-labelling the shape parameters in a corresponding set of images. Attachment of mentioned standard feature improvement techniques at the end of the pipeline, increases the efficiency and enhances the accuracy of the process. The approach is succeeded with the integration of a properly-designed DBN-HMM-based final recognizer system, which is explored by working on different DBN topologies and parameter settings. To the best of our knowledge, this feature extraction pipeline has not been devised in this way previously.

Experiments were conducted on the CUAVE corpus, in two separate tasks of the Multi-Speaker (MS) and Speaker Independent (SI). The proposed method was evaluated over both the phoneme- and word-level accuracies; feature extraction mechanism was formerly assessed with a baseline HMM recognizer.

The rest of this paper is organized as follows: section 2 reviews some recent seminal studies of lipreading. Section 3 describes the visual feature extraction process and Deep Neural Networks (DNNs), exploited for the classification. Section 4 reports the experimental results: first, the HMM baseline model with conventional visual features is explained. Next, various deep architectures of a certain neural network type, called DBN, are explored, and finally, phoneme- and word-level recognition rates are presented. Discussions are provided in Section 5, and conclusions are given in Section 6.

RELATED WORKS

Typically, there are some challenges with automating lip-reading process, including the coarticulation effect, visual features diversity and speaker-dependency of features. The history of these issues along with some proposed solutions for them could be found in the literature. Regarding to the coarticulation effect and visual feature diversity, the following studies have reached to noticeable solutions.

In (Lan *et al.*, 2010), it is demonstrated that the effects of inter-speaker variability of features could be reduced by applying the feature improvement of per speaker z-score normalization and the hierarchical LDA (HiLDA) (Potamianos *et al.*, 2004) techniques. Using HMM recognizer, the best viseme accuracies of 45% and 50% are achieved in the MS and SI tasks, respectively.

In (Lan *et al.*, 2012b) the best viewing angle for an automated lip-reading system is studied, using a purpose built audio-visual speech database called LiLiR. The database contains multi-camera, multi-angle recordings of a speaker reciting 200 sentences from the Resource Management Corpus, with the vocabulary size of 1000 words. Visual features extracted in this work are based on the Active Appearance Models (AAMs), including the shape, appearance, "CAT" (simple concatenation of the shape and appearance parameters), "CSAM" (application of PCA over mentioned concatenation), and HiLDA variations. These features are extracted from the respective views and are appended with their second derivatives (ΔΔ). For all features, z-score normalization is applied which has been shown that improves the separability of the features among involved classes.

Regarding to the speaker-dependency issue, several studies have been done for the single or MS conditions. Some of these studies propose deep architectures for their lip-reading systems. As it has been proven, the DNNs are effective tools for the feature extraction and classification tasks (Hinton *et al.*, 2012). A lot of researches are recently published in which the ASR systems are implemented by employing various deep learning techniques. Some of these works demonstrate the use of deep learning in the Audio-Visual Speech Recognition (AVSR) systems (Huang and Kingsbury, 2013, Mroueh *et al.*, 2015, Ngiam *et al.*, 2011, Noda *et al.*, 2014, Srivastava and

Salakhutdinov, 2012). However, the application of such techniques to the automatic lip-reading could be investigated more. In the following studies, different deep architecture lip-reading systems are reviewed and the rate of their speaker-dependency is highlighted.

In (Ngiam *et al.*, 2011), the PCA is applied to the mouth region-of-interest (ROI), and a deep autoencoder (AE) is trained to extract the bottleneck (BN) features. The features from the entire utterance were fed to a support vector machine ignoring the temporal dynamics of the speech. Digit classification accuracies of 64.4% and 68.7% are achieved in this study, using AVLetters and CUAVE databases, respectively.

In (Srivastava and Salakhutdinov, 2012), similar feature extraction approach proposed by (Ngiam *et al.*, 2011) is followed; this is done, while a deep Boltzmann Machine (DBM) is trained to extract multimodal representations. The employed features are extracted from the entire utterance and fed to a support vector machine. Working on CUAVE database, digit classification accuracy of 69% is achieved.

In (Noda *et al.*, 2014), a convolutional neural network (CNN) is proposed to act as the feature extraction block for a lip-reading system; the speaker's mouth area images along with the phoneme labels are used during the training phase. The evaluation is done on an audio-visual speech database comprising 300 Japanese words with six different speakers, each of which is modeled with an independent CNN. The average phone recognition rate (PRR) (for 40 phonemes, normalized with the number of samples for each phoneme over six speakers) of about 58% is attained in this paper, using 64×64 pixels of mouth area images as input.

In (Mroueh *et al.*, 2015), some deep multimodal learning methods are proposed to fuse speech and visual modalities in an AVSR system. Two unimodal deep networks are first trained separately over the audio and video data, and their final hidden layers are fused to form a joint feature which is further used in another deep network. The experiments are conducted on the IBM large vocabulary audio-visual studio database, leading to phone error rate (PER) of 69.36% and 41% in video-only and audio-only DNN, respectively.

In (Almajai *et al.*, 2016), a SI lip-reading system is designed, using Resource Management (RM) database. A combination of MLLT followed by SAT is applied over the CSAM features introduced in (Lan *et al.*, 2012b). HMM models are trained over phoneme and viseme units, but the phoneme implementation represents the superiority. The best word accuracies of 45% and 54% are achieved with HMM and DNN application, respectively.

In (Wand *et al.*, 2016), the lip-reading is succeeded with a processing pipeline, based purely on neural networks. In this study, Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) are stacked, so that a single structure is formed. Using the raw mouth images as the neural network inputs, the performance of such stacked network is experimentally evaluated and compared to the standard Support Vector Machine (SVM) classifier. Evaluations are performed in some speaker-dependent tasks on data from 19 speakers of GRID corpus (Cooke *et al.*, 2006). Averaging over the speakers, the best word recognition rate of 79.6% is reported, using end-to-end neural network-based architecture.

An end-to-end visual speech recognition system based on Long-Short Memory (LSTM) networks is proposed in (Petridis *et al.*, 2017). The model consists of two streams to extract features directly from the mouth and difference images, respectively. In this study, the temporal dynamics in each stream are modelled by an LSTM and the fusion of the two streams takes place via a Bidirectional LSTM (BLSTM). The classification accuracy of 84.5% over utterances is reported on the OuluVS2 database. This accuracy is reached to 78.6% on the CUAVE database with similar visual front-end.

In (Stafylakis and Tzimiropoulos, 2017), an end-to-end deep learning architecture for word level visual speech recognition is proposed. The proposed system is made of a spatiotemporal convolutional, residual and bidirectional Long Short-Term Memory networks. All evaluations are done on the Lipreading In-The-Wild benchmark, a challenging database of 500-size target-words consisting of 1.28sec video excerpts from BBC TV broadcasts. In this study, the word accuracy of 83.0% is attained by the proposed network.

In the current research, from the modeling perspective, the approach could be categorized as the neural network-based works reported in the literature. In particular, the final recognizer is designed based on the DBN-HMMs, implemented in the MS and SI tasks. The performance of the developed system is evaluated by the phone and word recognition accuracy measure, inside a digit recognition task. Working on a database of limit digits (0-9), makes it more rational to select phonemes as the recognized units. Thus, the PRRs are measured and reported primarily. Once a proper deep architecture DBN recognizer is explored, the WRRs are reported as well.

The effects of feeding the visual features to the involved DBN is evaluated by devoting a considerable part of this research to some proposed visual feature extraction schema.

MATERIAL AND METHODS

In this section, the proposed lip-reading method and its two major parts of feature extraction and classification are explained. First, video files are converted to the frame sequences, and the corresponding features are extracted. Next, some nearly optimized DBN architectures with proper parameter settings are sought and examined.

HYBRID VISUAL FEATURE EXTRACTION

Visual features have been broadly categorized as those depending on video pixels and those based on models; in the literature, these two kinds of features are referred to appearance- and shape-based features, respectively. Previous works have shown that the raw features, such as dimensionally reduced image intensities, can be used as the inputs to neural networks while resulting in good classifications (Ngiam *et al.*, 2011). In (Lan *et al.*, 2009) it is mentioned that if two types of appearance- and shape-based features, such as pixel intensities and contour shapes of the ROI, are combined, more informative feature vectors can be created. Accordingly, we devise our feature extraction process to have at least two function blocks of appearance- and shape-based feature extractors.

So, the lip-reading system is made up of the following function blocks: Face detection, lip localization followed by the ROI bounding box extraction, scaling and reshaping, meaningful feature extractions, techniques of inter-class discrimination and speaker adaptation inclusion, and dynamic information consideration. Many of these blocks are the same as used in conventional visual speech recognition systems; however, certain arrangement of those and, introduction of some different blocks make the whole process influence differently.

The general scheme of the proposed feature extraction process is represented in Fig. 1. As mentioned earlier, each video file is converted to a sequence of frames, in which the speaker's body is covered from the shoulders to the head. Each frame is preprocessed, so as to extract only the ROI encompassing the mouth. For this purpose, Active Shape Modeling algorithm (Cootes *et al.*, 1995) is used, via which face landmarks are obtained. Face detection is made by tracking the landmarks placed on the border of face. Eighteen landmarks are positioned on the inner and outer lip contours. Similarly, by tracking these points, an exact bounding box (B.B) of the lips is obtained.

Each determined mouth B.B is rescaled to 25×40 pixel size, and finally is converted to a 1000-dimensional feature vector. This vector is further transformed to a 32-dimensional vector, using the PCA algorithm. In this way, the so-called appearance-based features are created.

From the (x,y) lip coordinates, the width and height of the mouth are calculated and used as the geometric features. Concatenating the x- and y-coordinates with the geometric features, a 38-dimensional vector is created, which we call it the shape-based feature set. These features are combined with the appearance-based features, making a 70-dimensional feature vector.

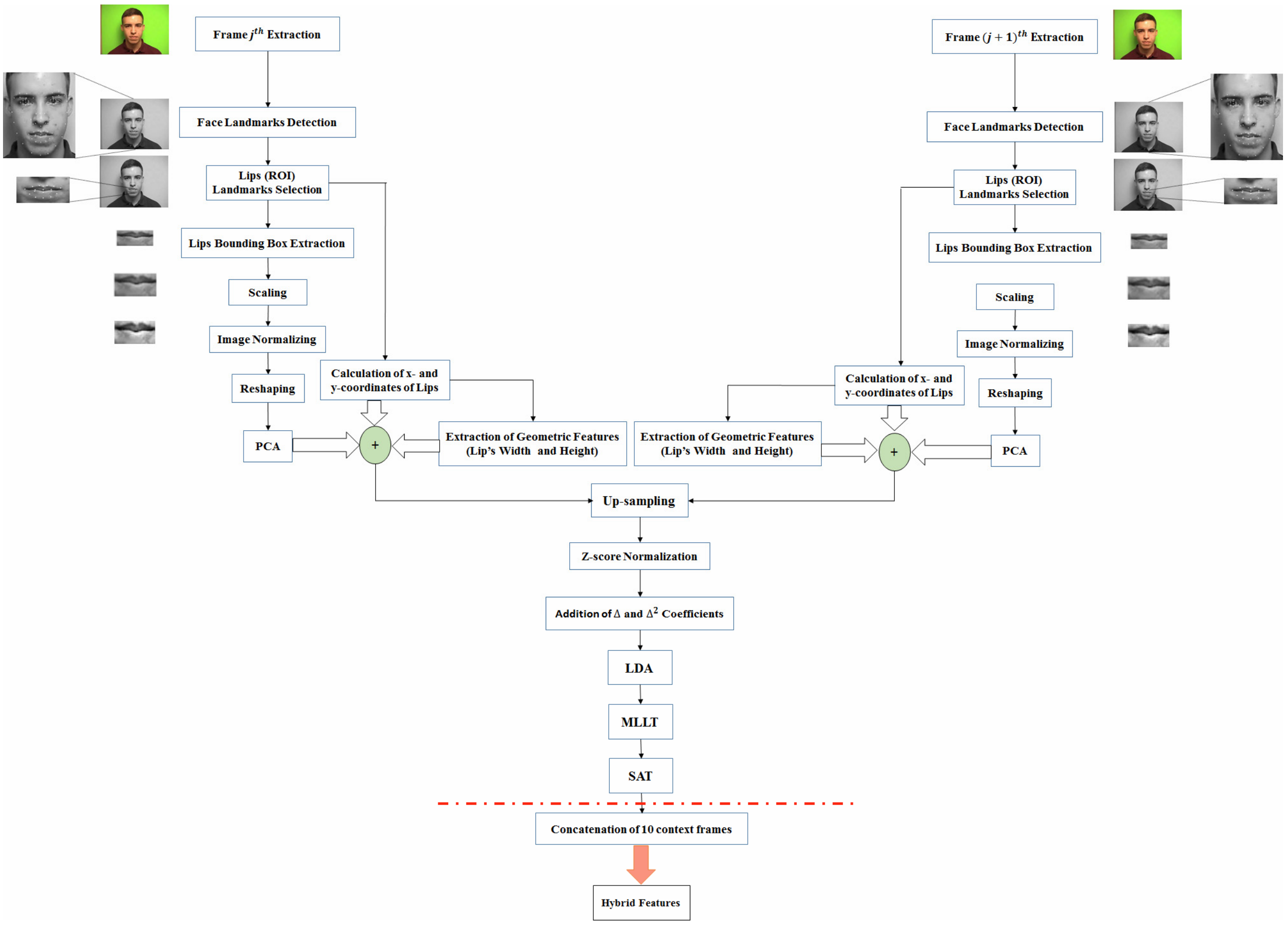
In CUAVE database, the video files are recorded with the NTST standard of 29.9 frames per seconds (fps) (Patterson *et al.*, 2002). These files are accompanied with audio files that are segmented into frames with the rate of 100 fps. The transcription is made according to the audio files, demonstrating 100 labels in one second. The sequence of the 70-dimensional vectors is then up-sampled to 100 fps, so that the direct mapping to the corresponding phoneme labels becomes possible.

To complete the feature vector created up to this stage, the z-score normalization is applied, and augmentation is made by adding the first and second derivatives (Δ/ΔΔ) of the evaluated features.

Extension of the process with the next three blocks of LDA, MLLT, and SAT is inspired from (Almajai *et al.*, 2016). In (Almajai *et al.*, 2016), it is shown that the MLLT and SAT techniques improve the phone and word recognition accuracies, significantly, especially when a DNN recognizer is applied thereafter. LDA is implemented to learn the optimal projection that best describes the dynamics of speech. From the so far acquired feature vector, it learns a set of orthogonal projections that maximize the between class distance and at the same time, minimize the within class distance. With LDA application, 40-dimensional feature vectors are resulted, which are further decorrelated using the standard technique of MLLT. With MLLT technique, a linear transform of the input features is found in which the assumption of a diagonal covariance matrix is the most valid. It can be shown that inter-class discrimination is improved in this condition. The SAT technique is then applied using feature-space Maximum Likelihood Linear Regression (fMLLR) of 40×41. With this technique, the effect of variation in the features of different speakers is normalized to model the intra-speaker variability and to avoid modelling of the inter-speaker variability. The 40-dimensional speaker adapted features are then spliced across a window of *n* frames, to make a hyper-vector, which we call it hybrid features in this paper. Contiguous frames are chosen to be (*n*–1)/2 frames before and after the central vector. How to choose the correct value for *n* will be discussed in "Multi-speaker task" section.

The PCA dimensionality reduction block was formerly used in conventional visual speech recognition systems; however, by deliberately implementing it after the blocks of detection, localization, scaling, normalization, and reshaping (represented in Fig. 1), it is aimed to achieve more beneficial appearance-based features to this point, while not facing the challenges of robustly tracking the lip-contour and hand-labelling the shape parameters (the *x* and *y*-coordinates of a set of *s* vertices that delineate the lip-contour) in a corresponding set of images. The greatest influence of the proposed feature extraction process comes from the augmentation of features with the higher level geometric features. Augmentation of the process with confirmed techniques of MLLT and SAT adds another weighty value to the whole process, which had not been considered in this manner before.

A dashed line in Fig. 1 is drawn to show where the HMM and DBN recognizers are differentiated in taking their input feature vectors (discussed in section "Baseline Model").



Frame j^{th} Extraction
Face Landmarks Detection
Lips (ROI) Landmarks Selection
Lips Bounding Box Extraction
Scaling
Image Normalizing
Reshaping
PCA
Calculation of x- and y-coordinates of Lips
+
Extraction of Geometric Features (Lip's Width and Height)
Frame $(j+1)^{th}$ Extraction
Face Landmarks Detection
Lips (ROI) Landmarks Selection
Lips Bounding Box Extraction
Scaling
Image Normalizing
Reshaping
PCA
Calculation of x- and y-coordinates of Lips
+
Extraction of Geometric Features (Lip's Width and Height)
Up-sampling
Z-score Normalization
Addition of Δ and Δ^2 Coefficients
LDA
MLLT
SAT
Concatenation of 10 context frames
Hybrid Features

Fig. 1. *The schematic of the hybrid feature extraction process.*

DEEP BELIEF NETWORK CLASSIFIER

Restricted Boltzmann Machine (RBM) is a stochastic neural network that is represented by a bipartite graph, where visible units are connected to hidden units using undirected weighted connections. The visible and hidden units can have any distribution in the exponential family (Welling *et al.*, 2005); however, the Bernoulli and Gaussian distributions are the mostly used distributions. The RBMs in a stack scheme can produce a single multi-layer generative model known as Deep Belief Network (Hinton *et al.*, 2006).

The weight initializing is one of the fundamental problems in the corresponding deep training process. Hinton *et al.* (Hinton *et al.*, 2006) introduced a layer-wise pre-training algorithm that led to a good initialization of the weights and biases. After pre-training, the DBNs are normally fine-tuned with respect to a typical supervised criterion such as the cross-entropy, to improve the network performance.

APPLICATION OF THE DBN IN THIS WORK

In order to apply DBN-HMM to the current phone recognition task, the proposed visual features are used to set the states of the visible units of the lower layer of the involved DBN; this produces a probability distribution over the possible labels of the central frame of the corresponding consecutive frames.

In the current work, the visual inputs are represented via real-valued feature vectors, and the hidden units are assumed to be binary. Therefore, the first building block of the used RBM is considered to have a Gaussian–Bernoulli (G–B) distribution; but, the other stacking RBMs have Bernoulli-Bernoulli (B–B) distributions.

The DBN-HMM system is trained using the alignment of states and the decision tree derived from the GMM stage. The mini-batch stochastic gradient descent technique is employed to initialize the weights of the DBN. For the involved G–B RBMs, 50 epochs are run with the learning rate of 0.05; while for the B–B RBMs, 25 epochs with a fixed learning rate of 0.6 are run. These learning rates are found to be optimal after testing different values. Next, fine-tuning stage is followed to complete the DNN-HMM hybrid model. To develop this hybrid model, DBNs with different architectures are trained to estimate the posterior probabilities of the corresponding HMM states (Veselý *et al.*, 2013). Hybrid models are then used as the implemented Visual Model (VM) to generate the VM scores for the lattice generating decoder.

Fig. 2 shows the overall architecture of the described lip-reading system. This architecture is used to decode the test video streams; this is done after the training phase in which the visual and language models (LMs) are trained over the selected training set. The Bigram LM is prepared from the constructed lexicon that affects weights of the lattice trees in the decoding process (Mohri *et al.*, 2008). All the implementations needed for realization of the mentioned classifier and conducting the experiments are accomplished by employing the Kaldi speech recognition toolkit (Povey *et al.*, 2011).

Feature extraction and deep learning based classification experiments are done on a consumer-class personal computer with an AMD FX(tm)-8350 Eight-Core processor (4 GHz, 4 cores), 32 GB RAM, and a single NVIDIA GeForce GTX Titan X graphic processing unit with 12 GB on-board graphics memory.

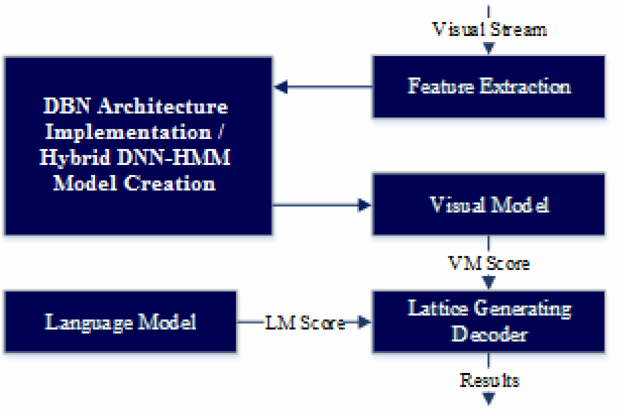

Visual Stream
Feature Extraction
DBN Architecture Implementation / Hybrid DNN-HMM Model Creation
Visual Model
VM Score
Language Model
LM Score
Lattice Generating Decoder
Results

Fig. 2. *Overall architecture of the implemented lip-reading system.*

To test the robustness of the proposed features across different speakers, two separate lip-reading tasks of the MS and SI are arranged. In the MS task, different DBN architectures are examined and the best model is selected. In the SI task, the investigation space is reduced to the best DBN architectures obtained in the previous task.

RESULTS

In this section, the employed database and a baseline model along with the conventional visual features for the automatic lip-reading purpose are presented. To have a more comprehensive comparison, the proposed hybrid features are also utilized as the input of the baseline model. Next, the experiments are conducted for the MS and SI tasks, and the corresponding results are reported alongside.

DATABASE

The experiments are conducted over the Clemson University Audio Visual Experiments (CUAVE) database (Patterson *et al.*, 2002). The main purpose of utilizing this corpus is to allow testing of more realistic speaker data that includes a variety of visual features and speaker movements. The corpus contains 36 individuals of different sex (17 female and 19 male), race and ethnicity. Speakers utter digits 0-9 in ascending and descending order in various styles of: standing still, moving side-to-side/back-and-forth/or tilting the head, turning to profile view, and then returning to the frontal view. Since the "frontal" view of lips seems to have enough information for lip-reading task, all "profile" views are neglected. In this way, two separate video parts are created out of each video file. Although, the moving sections of the videos are included in the experiments, the utterances with incomplete facial crops in the corresponding frames are also neglected due to their hardness of face recognition.

BASELINE MODEL

The Hidden Markov Model (HMM) is used as the baseline recognition models; this is the method of choice for speech recognition and have been shown to be successful for the automatic lip-reading purpose (Lan *et al.*, 2012a, Matthews *et al.*, 2002). In the current study, the HMMs are constructed so that a comparison could be made between the results obtained through the baseline system against the DBN-HMM deep architecture-based lip-reading system.

Kaldi toolkit is applied here for building and manipulating the employed HMMs. Since the performance of the developed system is evaluated for digit recognition task and the basic models are developed over phoneme units, a total of 19 separate phonemes and 1 extra code for the silence are considered and 20 HMMs are finally created. The built HMM system is a triphonic-based system considered as the baseline system. For each HMM, three states is considered, and the number of Gaussians per HMM state is dedicated automatically based on the data count (Povey *et al.*, 2011). To compose the Kaldi decoding graph, a phone-pair bigram language model is built and the lexicon is derived from the CUAVE database.

The proposed hybrid features along with the conventional Discrete Cosine Transform (DCT), Geometric, and ASM visual features are employed as the input of the baseline system. A feature selection process is made in a way to cover examples from different categories of the appearance-based, shape-based, and hybrid visual features.

To extract the ASM visual features, the Active Shape Model algorithm is used similar to what explained in "Materials and Methods" section. As mentioned earlier, frames of video files are up-sampled to 100 fps. Then, 77 landmarks are extracted via ASM algorithm, which represent (x,y) coordinates of face points. Out of all, 18 coordinates correspond to the points on the lips contour. These coordinates are gathered to make a 36-dimensional feature vector. Delta (Δ) and delta-delta (ΔΔ) coefficients are added to this vector, and z-score normalization is implemented on it. The resulted feature vector is called ASM-feature in this work.

To extract the Geometric visual features, a number of high level features that are meaningful to humans, including the contour height, width, perimeter, and the area within the contour, are extracted. In this regard, a box and a diamond encompassing ROIs are considered. Similar to the ASM-feature extraction procedure, the up-sampling, adding Δ/ΔΔ coefficients, and z-score normalizing are done, subsequently.

To observe the strength of transform domain features, DCT features are examined additionally. It is a useful transform for the dimensionality reduction purpose (Potamianos *et al.*, 2004). The area of lip is transformed into DCT coefficients after the application of the DCT. The ROI is transformed into 8×8 grayscale intensity image blocks; they are then converted to a single vector using zigzag strategy. In the literature, dimension of the DCT features has been chosen to be a value in the range of 20 to 35 (Morade and Patnaik, 2015, Petridis and Pantic, 2016, Potamianos and Neti, 2001). Accordingly, random dimensions of 24, 27, 30, and 33 are considered in this study. Experiments show that the best dimension among these values is 27. A DC (zero frequency) component is included in the features so that to provide information related to percentage visibility of teeth and tongue (Morade and Patnaik, 2015). Next, Δ/ΔΔ coefficients are added, and z-score normalization is performed, lastly. The resulted feature vector is called DCT-features in this work.

The proposed hybrid visual features are firstly fed to the baseline recognition system to have an initial evaluation of their strength. In such a condition, these hybrid features are employed without concatenating the context frames. This is done to make consistency with the size limitation of the conventional HMM input. The corresponding experimental results are given in Table 1 and Fig. 5. Fig. 5 represents the whole comparison of these features after applying the proposed features to the DBN-HMM system.

According to Table 1, using the hybrid features resulted in the PRR of 61.5%. This high accuracy demonstrates the strength of the proposed features, even when a conventional HMM based classifier is employed.

Table 1. *Comparison between the proposed hybrid features and traditional visual features, using baseline HMM classifier.*

Category	Visual Feature Type	PRR (%)
Appearance-based	DCT	38.9
Shape-based	Geometric	40.9
	ASM	38.65
Hybrid	Proposed Hybrid Features	**61.5**

MULTI-SPEAKER TASK

To handle the previously defined tasks, various DBN topologies are examined in this section. The MS experiments are designed in a way that all the speakers attend in the training phase. For this purpose, a special version of the 6-fold cross-validation technique is used.

As mentioned earlier, the employed database is conditioned in a way to include two distinct video parts for each speaker. Our plan is to divide the individual sections of the database into three subdivisions of 66.66% (24 speakers), 16.66% (6 speakers) and 16.66% (6 speakers) to devote to training, testing and development sets, respectively. However, to satisfy the condition of all speakers attendance during the training, both video parts of the first 12 speakers, along with the first video parts of the remaining 24 speakers (12 of the train, 6 of the development, and 6 of the test subdivisions) are used to train the DBN; while, second video parts of 24 remained speakers are put for the test and development sets. This technique is then repeated for selecting all the 6 possible held up group.

Using hybrid visual features

Here, the visual features that are proposed in “Materials and Methods section” are used as the DBN inputs, while different DBN topologies are examined. The initial DBN architecture is inspired from (Mohamed *et al.*, 2012) and defined as 440 – 1024 – 1024 – 1024 – 1024 – 60; where 440 is the network input dimension spliced across a window of 11 frames, 1024 is the number of intermediate hidden units, and 60 is the network output dimension, corresponded to the HMM states. As mentioned before, each phoneme class is modeled with a single HMM of 3 states. Thus, in overall 60 states are resulted for all phonemes.

This architecture is modified parametrically in four separate experiments: first, the effect of varying the number of hidden units (network’s width) is studied. In this regard, 256, 512, 1024 and 2048 units are distinctively examined for the intermediate RBMs. Secondly, the effect of the number of hidden layers (network’s depth) is examined. Accordingly, different layer numbers of 2, 3, 4, 5, 6, and 7 are considered. Thirdly, the effect of context frame numbers is investigated. So, a wide range of window frames, from 3 to 27, are considered. Finally, the effect of learning rate parameter has been investigated. Different pair numbers, corresponding to the G–B and B–B RBMs, is considered, respectively. Resulted accuracies are represented in Fig. 3.

According to the results represented in Fig. 3, a DBN network with a width of 1024, depth of 4 hidden layers (6 layers in overall), and 11 context frames in the input could result in a PRR of 71.65%, when the learning rate is set to 0.4 and 0.01, during the train of the G–B and B–B RBMs, respectively. However, by increasing the input window frames to 21, the PRR is improved from 71.65% to 75.1%. By changing the learning rates to 0.6 and 0.05, for the G–B and B–B RBMs, respectively, this recognition accuracy is further improved to 77.65%, which is significant. As a result, the best examined DBN architecture is found to be 840 – 1024 – 1024 – 1024 – 1024 – 60, which leads to the PRR of 77.65%.

It is inferred that (a) the same size for every hidden layer in the network could be considered while still achieve appropriate accuracies. This size is found to be 1024 in the current study, (b) 1024 neurons are able to code the higher layer visual representations, when the task is phone recognition. This means that 1024 neurons are able to cover visual information of the lips when the speaker utters phonemes, (c) More hidden layers improves performance when the number of hidden layers is less than 4, but with more hidden layers, the accuracy begins to descend. The complexity of the visual speech recognition could be handled with a network of 4 hidden layers, (d) a window of 21 frames (equivalent to 525ms) in this task, covers the average size of phones and its required dynamic information. Smaller input windows miss important discriminative information in the context, while networks with larger windows are probably getting distracted by the almost irrelevant information far from the center of the window, and (f) learning rates play significant roles that should not be neglected.

For the resulted proper DBN topology, the WRR is found to be 80.25%, which is noteworthy.

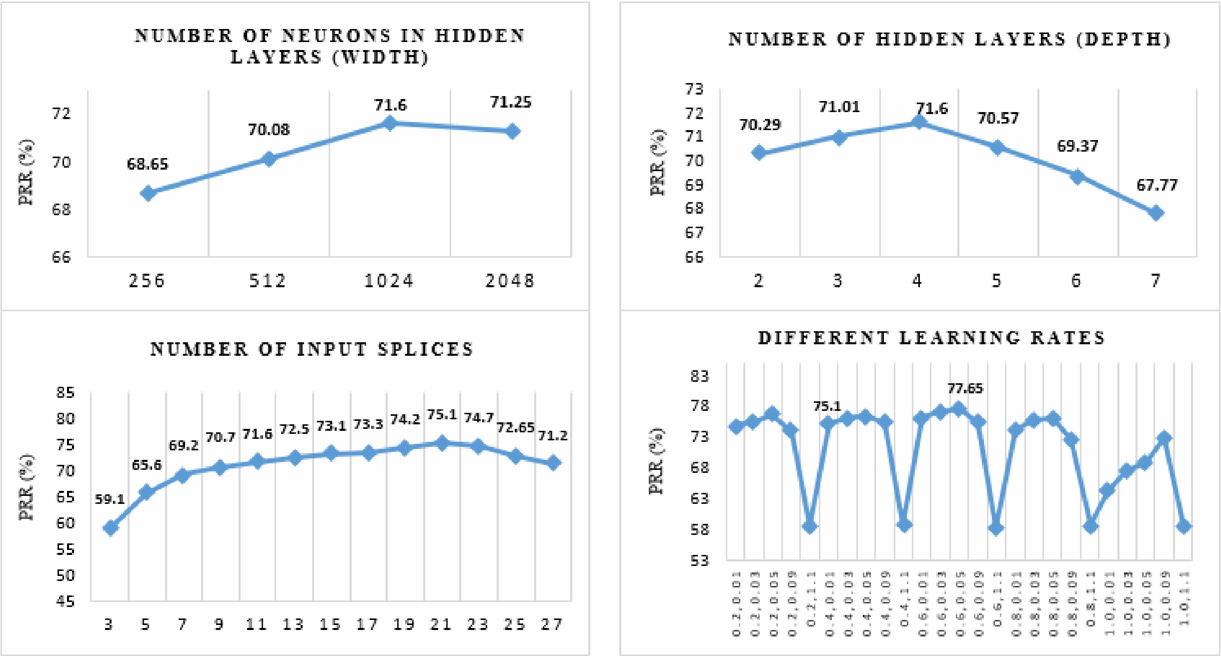
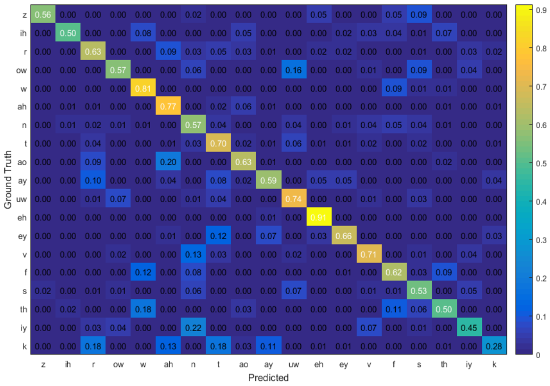


Fig. 3. *From left to right, along rows: the effect of DBN's width, depth, splice numbers, and learning rate on PRR.*

SPEAKER-INDEPENDENT TASK

In the SI task, only the best architecture found in the previous experiments (840 – 1024 – 1024 – 1024 – 1024 – 60) is considered for further investigations. As mentioned in the MS task, the individual sections of the database are divided into three parts of 66.66%, 16.66% and 16.66% to devote to training, testing and development sets, respectively. A 6-fold cross-validation experiment is considered, where for each fold, six speakers are present that were absent during the training and evaluation phases. In each fold, the whole utterances of those 6 speakers are used for testing, while the utterances of the rest of 30 speakers are used for training and development processes. This technique is then repeated for all the 6 possible held up groups. Using the 840 – 1024 – 1024 – 1024 – 1024 – 60 architecture, the SI PRR of 73.40% is achieved. The corresponding WRR is measured as well and found to be 76.91%, which is remarkable.

The resulted accuracies reveal that the involved deep architecture is designed properly that is able to act appropriately even in the challenging SI task, in which large variation in lip-shapes of speakers is clearly seen. As expected, here, the confusions are higher, and the rate of misclassifying is raised. However, the confusion matrix (Fig. 4) shows that the overall classification is acceptable for this challenging task. Reading the confusion matrix from left to right, each entry is the number of times a phoneme class is visually recognized as another one. The entries are normalized and expressed as a percentage to the whole samples of every targeted phoneme. Row and column of the "silence" are removed for clarity.



	z	ih	r	ow	w	ah	n	t	ao	ay	uw	eh	ey	v	f	s	th	iy	k
z	0.56	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.05	0.00	0.00	0.05	0.09	0.00	0.00	0.00
ih	0.00	0.50	0.00	0.00	0.08	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.02	0.03	0.04	0.01	0.07	0.00	0.00
r	0.00	0.00	0.63	0.00	0.00	0.09	0.03	0.05	0.03	0.01	0.00	0.02	0.02	0.00	0.00	0.01	0.00	0.03	0.02
ow	0.00	0.00	0.00	0.57	0.00	0.00	0.06	0.00	0.00	0.00	0.16	0.00	0.00	0.01	0.00	0.09	0.00	0.04	0.00
w	0.00	0.00	0.00	0.00	0.81	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.09	0.01	0.01	0.00	0.00
ah	0.00	0.01	0.01	0.00	0.00	0.77	0.00	0.02	0.06	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.01
n	0.00	0.01	0.02	0.01	0.01	0.00	0.57	0.04	0.00	0.00	0.04	0.00	0.01	0.04	0.05	0.04	0.01	0.00	0.00
t	0.00	0.00	0.04	0.00	0.00	0.01	0.03	0.70	0.02	0.01	0.06	0.01	0.01	0.02	0.00	0.02	0.00	0.00	0.01
ao	0.00	0.00	0.09	0.00	0.00	0.20	0.00	0.00	0.63	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00
ay	0.00	0.00	0.10	0.00	0.00	0.04	0.00	0.08	0.02	0.59	0.00	0.05	0.05	0.00	0.00	0.00	0.00	0.00	0.04
uw	0.00	0.00	0.01	0.07	0.00	0.00	0.01	0.04	0.00	0.00	0.74	0.00	0.00	0.01	0.00	0.03	0.00	0.00	0.00
eh	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.91	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ey	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.12	0.00	0.07	0.00	0.03	0.66	0.00	0.00	0.00	0.00	0.00	0.03
v	0.00	0.00	0.00	0.02	0.00	0.00	0.13	0.03	0.00	0.00	0.02	0.00	0.00	0.71	0.00	0.01	0.00	0.04	0.00
f	0.00	0.00	0.00	0.00	0.12	0.00	0.08	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.62	0.03	0.09	0.00	0.00
s	0.02	0.00	0.01	0.01	0.00	0.00	0.06	0.00	0.00	0.00	0.07	0.00	0.00	0.01	0.01	0.53	0.00	0.05	0.00
th	0.00	0.02	0.00	0.00	0.18	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.11	0.06	0.50	0.00	0.00
iy	0.00	0.00	0.03	0.04	0.00	0.00	0.22	0.00	0.00	0.00	0.00	0.00	0.00	0.07	0.00	0.01	0.00	0.45	0.00
k	0.00	0.00	0.18	0.00	0.00	0.13	0.00	0.18	0.03	0.11	0.00	0.01	0.01	0.00	0.00	0.00	0.01	0.00	0.28

Fig. 4. *Confusion matrix for the test data, produced by applying the best DBN architecture to the SI task.*

As demonstrated, almost all misclassifying phonemes are less than correctly classified ones. Vowels (*e.g.*, /*ah*/, /*ow*/, /*ao*/, /*ay*/, /*uw*/) are relatively well discriminated and recognized. It is attributed to the fact that the generation of vowels is strongly correlated with the visual cues represented by the movements of lips or jaw (Barker and Berthommier, 1999, Yehia *et al.*, 1998).

However, some consonants have been less efficiently recognized (*e.g.* /*k*/, /*s*/, /*z*/, /*th*/). The articulation of these consonants is mostly attributed to the dynamic interaction of the interior oral structures, such as tongue, teeth, oral cavity, which are invisible from the frontal facial images (Noda *et al.*, 2014). This is not the case in recognition of *'v'* and *'w'*; these are consonants that are mostly attributed to the lips and are visually tractable. The *'r'* consonant is not labial, but is obviously palatal, hence, could be mostly distinguished visually. As represented in Fig. 4, the correct rate of classification of these 3 consonants are interestingly high.

DISCUSSION

In this paper, some issues of the automatic lip-reading systems are discussed and accordingly in two aspects the following cases are proposed and developed: (a) a modified method of extracting appropriate visual features (b) a properly designed Deep Belief Network structure accommodated inside the hybrid structure implemented by the Kaldi toolbox.

Applying only the feature extraction modifications, the proposed hybrid visual features are utilized as the input of the baseline HMM and higher PRR (61.5% vs. the topmost acc. of 40.9%) are resulted. This accuracy is significantly increased (to 77.65%), when the properly structured DBN-HMM recognizer is replaced. Details of these comparisons are represen-ted in Fig. 5.

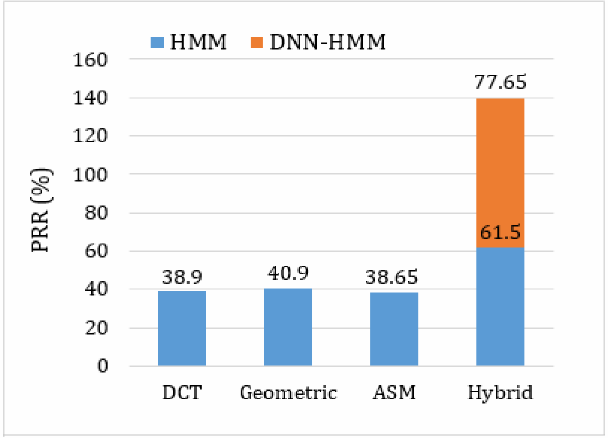


Fig. 5. *The PRRs obtained in MS task, using proposed and traditional visual features in HMM and DNN-HMM recognizers.*

Some recognition accuracies reported in the literature in the visual speech recognition/lip-reading field are scheduled in Table 2. Task comprehensiveness is highlighted and shown in yellow color: some studies applied their developed systems distinctively over both the MS and SI tasks. The superiority of model strength and its robustness could be concluded inside the table. In the last column, the accuracies are reported based on the determined classification level.

Table 2. *Review of: seminal studies about visual speech recognition/ lip-reading, their highlights, and accuracies.* $^{++}$ *these accuracies are achieved by incorporation of sophisticated deep recognizers.*

Study	Corpus	Feature	Method	Task		Class Labels			Best Accuracy (%)
				MS	SI	Phoneme	Viseme	Word	
(Lan *et al.*, 2010)	LiLiR	HiLDA projected AAM + z-score normalization	HMM	+	+		+		45 (the MS task) 50 (the SI task)
(Ngiam *et al.*, 2011)	CUAVE	ROI + PCA + Δ/ΔΔ + Context frames	Deep AE + SVM		+			+	68.7
(Lan *et al.*, 2012a)	LiLiR	Cat + ΔΔ	HMM		+		+		52.5
(Srivastava and Salakhutdinov, 2012)	CUAVE	ROI + PCA + Δ & ΔΔ + Context frames	Multi DBM + SVM		+			+	69
(Noda *et al.*, 2014)	Japanese database	Raw ROIs	CNN	+		+		+	48 (PRR) 22.5 (WRR)
(Mroueh *et al.*, 2015)	IBM AV-ASR	ROI + LDA + Context frames	DNN		+	+			30.64
(Almajai *et al.*, 2016)	RM	scam + LDA + MLLT + SAT	HMM		+			+	45
			DNN						54
(Wand *et al.*, 2016)	GRID	Raw ROIs	LSTM	+				+	79.6^{++}
(Petridis *et al.* 2017)	CUAVE	Raw ROIs	LSTM		+			+	78.6^{++}
	OuluVS2								84.5^{++}
(Stafylakis and Tzimiropoulos, 2017)	In-The-Wild	Face ROI + resizing + normalizing	CNN + ResNet + BLSTM		+			+	83.0^{++}
This work	CUAVE	Hybrid features	HMM	+	+	+		+	**61.5 (PRR)**
			DBN-HMM						**77.65 (the MS PRR)** **80.25 (the MS WRR)** **73.40 (the SI PRR)** **76.91 (the SI WRR)**

By considering the highlights and challenges, the discussion is followed in two branches: comparing the current study results with those: (a) worked on the CUAVE database, and (b) worked on different database.

Studies that worked on the CUAVE database are shown in blue color in Table 2. Since there is not a standard partitioning for this database, it is difficult to exactly compare between reported results of different works.

In (Ngiam *et al.*, 2011), a simple visual feature extraction front-end is introduced along with new concepts regarding to multimodal audio-visual feature extraction. A word recognition rate of 68.7% was reported using video deep autoencoder (AE), followed by the SVM classifier. In (Ngiam *et al.*, 2011), the idea of feature learning when two sources of audio and video information are available is proposed. Bottleneck (BN) features were extracted that code information of audio, video, and their correlations, simultaneously. In other words, to create BN features, it could be interpreted that the audio information was used primarily and the visual features played the auxiliary role. However, in the current study, the hybrid features are extracted from the video only channel. Although, it was expected that the BN representations employed in (Ngiam *et al.*, 2011) were more beneficial for the ultimate recognition tasks, the WRR (76.91%) reported in the SI task of the current study reveals the strength of the proposed hybrid visual features.

In (Srivastava and Salakhutdinov, 2012), the similar approach proposed in (Ngiam *et al.*, 2011) was followed, except that a new architecture of multimodal DBM was introduced; as the result, the word recognition rate of 69% was reported. Again, although the extracted multimodal features in (Srivastava and Salakhutdinov, 2012) encoded both the audio and video information, the solely video encoding performance of the proposed hybrid visual features in the current task seems to be superior.

In (Petridis *et al.*, 2017), an end-to-end recognizer is introduced based on LSTMs, which are proven to have great performance in Classification tasks. The similarity of (Petridis *et al.*, 2017) and the current study is in the utilized CUAVE database. Over this database, the best reported SI WRRs are 78.6% and 76.91%, respectively. The 2% increase in the word recognition accuracy is definitely rational due to more complicated recognizer implementation.

In the following, comparisons with the rest of studies will be made. As stated earlier, these studies employed different databases.

The best SI PRR (73.40%) obtained in the current study could be compared with the viseme accuracies of the SI lip-reading systems reported in (Lan *et al.*, 2012a, Lan *et al.*, 2010). These systems utilized the AAM-based features in their front-end recognizers and resulted in viseme accuracies of 50% and 52.5%, respectively. Although LiLiR vocabulary is larger than CUAVE and hence the phoneme size is greater (44 vs 19), the far larger speakers population (about threefold) in the current study, reveals the strength of the proposed hybrid visual features.

This superiority could be also seen against the video-only DNN model proposed in (Mroueh *et al.*, 2015). A PER of 69.36% (accuracy of 30.64%) was achieved in this study, using IBM large vocabulary audio-visual studio database. Of course, part of this 42.76% difference between the obtained accuracies (30.64% vs 73.40%) is definitely due to the number of lexicons and phonemes in IBM vs CUAVE databases.

In (Almajai *et al.*, 2016, Lan *et al.*, 2010, Mroueh *et al.*, 2015), and this study, essentially viseme or phoneme class labels are recognized. So, the corresponding reported accuracies are rather more comparative. The comparisons are represented in Fig. 6.

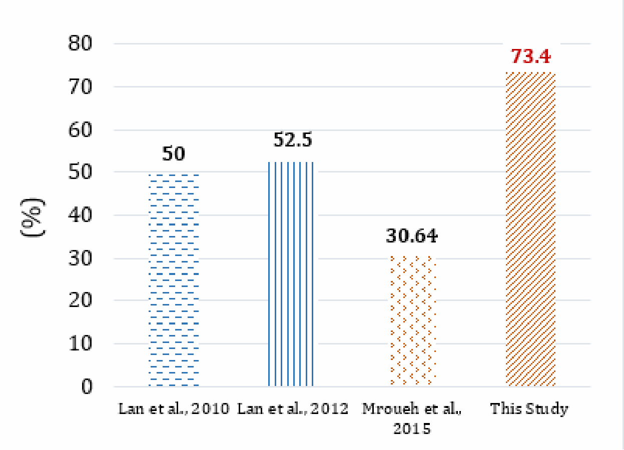


Fig. 6. *Viseme/Phone recognition rates, reported by some reviewed studies. Blue charts represent viseme recognition rates, while red charts depict phone recognition rate.*

The best MS PRR obtained in the current study (77.65%) is higher than the corresponding PRR of 48%, reported in (Noda *et al.*, 2014). Noda *et al.* considered speaker-dependent settings and utilized speaker-wise CNN models. Although Japanese database vocabulary and phoneme sizes are larger than those of CUAVE (300 vs 10 and 40 vs 19, respectively), achieving 77.65% accuracy is noteworthy, since higher recognition rate is resulted and more speakers (36) are incorporated.

Feature extraction and classification parts of the current study and (Almajai *et al.*, 2016) are comparable since in both engineered features are extracted, standard techniques such as MLLT and SAT are used, and the DNN-based classifiers are utilized. The superiority of the current study over (Almajai *et al.*, 2016) could be accounted for more speakers (about twofold) involvement. However, larger lexicon size (1000 vs. 10 words) and greater phoneme numbers (44 vs 19) of the RM over the CUAVE database compensates this superiority. The WRR reported in the current study is higher (76.91% vs. 54%) than the what reported in (Almajai *et al.*, 2016). It reveals that the beginner blocks (up to LDA) of the proposed feature extraction pipeline play effective roles that can lead to higher WRR. Thus, the arrangement of blocks meets the suitability.

The current study and (Wand *et al.*, 2016) have few in common: the databases, tasks, and classification levels are different. Besides, the MS and speaker-wise conditions are considered in this study. The fact is that in (Wand *et al.*, 2016) much more context frames are considered by the LSTMs, and hence the effect of dynamic information is highlighted. In the current study, however, the emphasis was on the feature extraction part, and the scope of view was reduced to this end.

Similarly, the current study and (Stafylakis and Tzimiropoulos, 2017) are hardly comparable. It is due to their completely different deep recognizers. In the current study, much simpler deep recognizer is utilized, while in (Stafylakis and Tzimiropoulos, 2017), a sophisticated network made up of a spatiotemporal convolutional, residual and bidirectional Long Short-Term Memory is incorporated. The results of the (Stafylakis and Tzimiropoulos, 2017) is therefore, expected to be higher since the network is devised more complicated. Besides, the lexicon size and type of databases are different too. As mentioned earlier, the focus of this study was mainly on the proper feature extraction, which are claimed to be strength enough if utilized in other recognizers as well.

CONSLUSION

This study has discussed some issues of the automatic lip-reading systems and explored an overcoming approach with two major parts of visual feature extraction and classification.

In the feature extraction part, hybrid features were extracted in a proposed well-designed arrangement of certain function blocks. As the final classifier, a single DBN with a well-designed topology, inside a DBN-HMM structure, was established and employed for all speakers. Experiments were applied on the CUAVE database, in two separate tasks of the MS and SI conditions. The proposed method was basically evaluated on the phoneme-level and considerably outperformed the HMM baseline recognizer.

The proposed visual feature set, so-called hybrid features, was used in two different lip-reading systems. Using the conventional HMM based recognizer as the baseline system, the PRR of about 61.5% was obtained. Introducing the features to the input of a DBN-HMM recognizer, with a well-designed DBN topology, led to the PRR of 77.65% for the MS task, which showed a considerable jump against the baseline system. In the SI task, the best PRR of 73.40% was achieved, which was highly appreciable, since the SI task has been proven to be challenging (Petridis *et al.*, 2017).

The PRR and WRR achieved in this study, were comparable with the corresponding accuracies reported in seminal recent literatures that were reviewed in this paper. It is verified that the proposed hybrid features are discriminative enough to be efficiently used in this task. As a future work, proposed hybrid feature set could be examined inside an Audio-Visual Automatic Speech Recognition (AV-ASR) system to investigate its benefits in combination with the audio based features.

REFERENCES

Almajai I, Cox S, Harvey R, Lan Y (2016). Improved speaker independent lip reading using speaker adaptive training and deep neural networks. 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 272–6.

Barker JP, Berthommier F (1999). Evidence of correlation between acoustic and visual features of speech. Ohala et al:199–202.

Bowden R, Cox S, Harvey RW, Lan Y, Ong E-J, Owen G, Theobald B-J (2012). Is automated conversion of video to text a reality? Optics and Photonics for Counterterrorism, Crime Fighting, and Defence VIII, volume SPIE 8546:85460U.

Cooke M, Barker J, Cunningham S, Shao X (2006). An audio-visual corpus for speech perception and automatic speech recognition. The Journal of the Acoustical Society of America 120:2421–4.

Cootes TF, Taylor CJ, Cooper DH, Graham J (1995). Active shape models-their training and application. Computer Vision and Image Understanding 61:38–59.

Hinton G, Deng L, Yu D, Dahl GE, Mohamed A-r, Jaitly N, Senior A, Vanhoucke V, Nguyen P, Sainath TN

(2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal Processing Magazine 29:82–97.

Hinton GE, Osindero S, Teh Y-W (2006). A fast learning algorithm for deep belief nets. Neural Computation 18:1527–54.

Hochreiter S, Schmidhuber J (1997). Long short-term memory. Neural Computation 9:1735–80.

Huang J, Kingsbury B (2013). Audio-visual deep learning for noise robust speech recognition. 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 7596–9.

Lan Y, Harvey R, Theobald BJ, Ong EJ, Bowden R (2009). Comparing visual features for lipreading. 2009 International Conference on Auditory-Visual Speech Processing, 102–6.

Lan Y, Theobald BJ, Harvey R, Ong EJ, Bowden R (2010). Improving visual features for lip-reading. Prodedings of the 2010 Conference on Auditory-Visual Speech Processing.

Lan Y, Harvey R, Theobald BJ (2012). Insights into machine lip reading. 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 4825–8.

Lan Y, Theobald BJ, Harvey R (2012). View independent computer lip-reading. 2012 IEEE International Conference on Multimedia and Expo (ICME), 432–7.

Matthews I, Cootes TF, Bangham JA, Cox S, Harvey R (2002). Extraction of visual features for lipreading. IEEE Transactions on Pattern Analysis and Machine Intelligence 24:198–213.

McClain M, Brady K, Brandstein M, Quatieri T (2004). Automated lip-reading for improved speech intelligibility. Proceedings of the 2004 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 1:I–701.

Mohamed A-r, Dahl GE, Hinton G (2012). Acoustic modeling using deep belief networks. IEEE Transactions on Audio, Speech, and Language Processing 20:14–22.

Mohri M, Pereira F, Riley M (2008). Speech recognition with weighted finite-state transducers. Springer handbook of speech processing 559–84.

Mroueh Y, Marcheret E, Goel V (2015). Deep multimodal learning for audio-visual speech recognition. 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2130–4.

Ngiam J, Khosla A, Kim M, Nam J, Lee H, Ng AY (2011). Multimodal deep learning. Proceedings of the 28th International Conference on Machine Learning (ICML-11), 689–96.

Noda K, Yamaguchi Y, Nakadai K, Okuno HG, Ogata T (2014). Lipreading using convolutional neural network. 15th Annual Conference of the International Speech Communication Association.

Patterson EK, Gurbuz S, Tufekci Z, Gowdy JN (2002). Cuave: A new audio-visual database for multimodal human-computer interface research. 2002 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2:II-2017.

Petridis S, Pantic M (2016). Deep complementary bottleneck features for visual speech recognition. 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2304–8.

Petridis S, Li Z, Pantic M (2017). End-to-end visual speech recognition with lstms. 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2592-6.

Potamianos G, Neti C (2001). Improved roi and within frame discriminant features for lipreading. Proceedings of the 2001 International Conference on Image Processing, 3:250–3.

Potamianos G, Neti C, Gravier G, Garg A, Senior AW (2003). Recent advances in the automatic recognition of audiovisual speech. Proceedings of the IEEE 91:1306–26.

Potamianos G, Neti C, Luettin J, Matthews I (2004). Audiovisual automatic speech recognition: An overview. Issues in Visual and Audio-Visual Speech Processing 22:23.

Povey D, Ghoshal A, Boulianne G, Burget L, Glembek O, Goel N, Hannemann M, Motlicek P, Qian Y, Schwarz P, Silovsky J (2011). The kaldi speech recognition toolkit. IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, EPFL-CONF-192584.

Savchenko A, Khokhlova YI (2014). About neural-network algorithms application in viseme classification problem with face video in audiovisual speech recognition systems. Optical Memory and Neural Networks 23:34–42.

Srivastava N, Salakhutdinov RR, (2014). Multimodal learning with deep boltzmann machines. Advances in Neural Information Processing Systems, 2222–30.

Stafylakis T, Tzimiropoulos G (2017). Combining residual networks with lstms for lipreading. arXiv preprint arXiv:1703.04105.

Sumby WH, Pollack I (1954). Visual contribution to speech intelligibility in noise. The Journal of the Acoustical Society of America 26:212–5.

Veselý K, Ghoshal A, Burget L, Povey D (2013). Sequence-discriminative training of deep neural networks. Interspeech 2345–9.

Wand M, Koutník J, Schmidhuber J, (2016). Lipreading with long short-term memory. 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 6115–9.

Welling M, Rosen-Zvi M, Hinton GE, (2005). Exponential family harmoniums with an application to information retrieval. Advances in Neural Information Processing Systems, 17:1481–8.

Yehia H, Rubin P, Vatikiotis-Bateson E (1998). Quantitative association of vocal-tract and facial behavior. Speech Communication 26:23–43.