Lip-reading via a DNN-HMM Hybrid System Using Combination of The Image-based and Model-based Features

Mohammad Hasan Rahmani[1], Farshad Almasganj[2*]
Biomedical Engineering Department, Amirkabir University of Technology (Tehran Polytechnic)
Tehran, Iran
mhnrahmani@aut.ac.ir, almas@aut.ac.ir

***Abstract*— Introducing features that better represent the visual information of speakers during the speech production is still an open issue that highly affects the quality of the lip-reading and Audio Visual Speech Recognition (AVSR) tasks. In this paper, three different types of visual features from both the image-based and model-based ones are investigated inside a professional lip reading task. The simple raw gray level information of the lips Region of Interest (ROI), the geometric representation of lips shape and the Deep Bottle-neck Features (DBNFs) extracted from a 6-layer Deep Auto-encoder Neural Network (DANN) are three valuable feature sets compared while employed for the lip reading purpose. Two different recognition systems, including the conventional GMM-HMM and the state-of-the-art DNN-HMM hybrid, are utilized to perform an isolated and connected digit recognition task. The results indicate that the high level information extracted from deep layers of the lips ROI can represent the visual modality with advantage of "high amount of information in a low dimension feature vector". Moreover, the DBNFs showed a relative improvement with an average of 15.4% in comparison to the shape features and the shape features showed a relative improvement with an average of 20.4% in comparison to the ROI features over the test data.**

Keywords—lip-reading; feature extraction; deep auto-encoder; DBNF

I. INTRODUCTION

Use of visual information in speech recognition, called lip-reading, is widely employed by people with hearing disabilities. All the information that a normal person commonly gets by the ears is understood only by the eyes in a deaf. Additionally, people with normal hearing would look at the speaker when the audio source is not enough to apperceive the spoken information in noisy environments. The main application of the lip-reading in known artificial intelligence tasks may be found in the Audio Visual Speech Recognition (AVSR) mission [1-3].

Selection of proper visual features, needed for the automatic lip-reading purpose, that best describes the visual information in a brief dimension size is not completely resolved [2, 4] until now. Due to the high role of selecting proper visual features in improving performance of the lip-reading process, this paper mainly focuses over comparison of three different types of visual features in a spoken visual digit recognition task.

Based on a definition introduced in [4], visual features can be classified into image-based and model-based types. The image-based features consist of crude gray level images taken from the video frames that mainly include lips area. This kind of features may be either used directly or after some processing. On the other side, the model-based features represent geometric concepts of lips appearance. Although the image-based feature sets contain much more information about the light intensity, teeth and tongue situation and many other high level information, the dimension of feature vector in the model-based systems is often less than the image-based ones resulting in an easier speech modeling and somewhere, higher recognition accuracy.

In this paper, lips area is extracted from video frames as the Region of Interest (ROI) and employed as a pure image-based feature set directly in the implemented lip-reading task. Another type of feature vector used in this paper is the shape of the lips represented by some fixed points on the inner and outer lips contour. Moreover, some high level deep components of the lips ROI are extracted via a nonlinear dimension reduction process, using Deep Auto-encoder Neural Networks (DANN). The output of the middle layer of such network, called Deep Bottle-neck Features (DBNFs), is used as a low dimension feature vector in the current lip-reading task. The obtained results show that DBNFs extracted from the pure image-based features are more beneficial in comparison to only the lips shape features. A hybrid of the Deep Neural Network and Hidden Markov Model (DNN-HMM) is employed to model the visual speech sequences and all the training and decoding processes is done on CUAVE [5] database of audio-visual spoken digits.

II. MATERIALS AND METHODS

A. Feature extraction

1) Raw gray level ROI features

The simplest feature set used in this work is the raw image of lips ROI in the gray level format. In order to evaluate these features, a face recognition process is first applied to each frame of the video streams to find the face bounding box. Another recognition process is then performed on the face

978-1-5090-6454-0/17/$31.00© IEEE

rectangle to find the appropriate bounding box that includes speaker's lips. Next, cropped image of lips area, extracted from the RGB-format frame is converted to the grayscale format and resized to a 25×40-pixel image. The resizing step is done to make all feature vectors meet the same size, that is necessary in modeling and recognition steps. Finally, the resultant matrix is converted to a 1000×1 vector by concatenating the columns of the grayscale image. Fig. 1 shows the procedure of preparing the gray level ROI features.

2) Lips shape features

Features extracted from the lips shape are widely used in the lip-reading and AVSR projects, commonly along with the appearance of visual elements of human speech production system [2-4, 6, 7]. Here, actually, some beneficial geometric information about the speakers' lips that describe the lips temporary shape briefly, are evaluated.

The Active Shape Model (ASM) approach [8] is employed to locate landmarks on speaker's lips, using STASM library [9]. In order to make the feature vector, the coordinates of all the lips inner and outer landmarks must be concatenated. Furthermore, the height and width of the lips are evaluated and added to the feature vector. Fig. 2a shows an example of the 18 landmarks of the lips, and Fig. 2b shows the height and width that are calculated from the corresponding landmarks and included in the final feature vector.

3) Deep bottleneck features

The high dimension of vectors in the raw lips ROI features and the relatively reduction of information occurring in the shape based features led us to propose another type of features that include almost all the information of raw lips ROI in a low dimension of the generated feature vector. For this purpose, a 6-layer DANN that reduces the 1000-dimensional lips ROI to a 10-element brief high level vector, called DBNFs, is exploited.

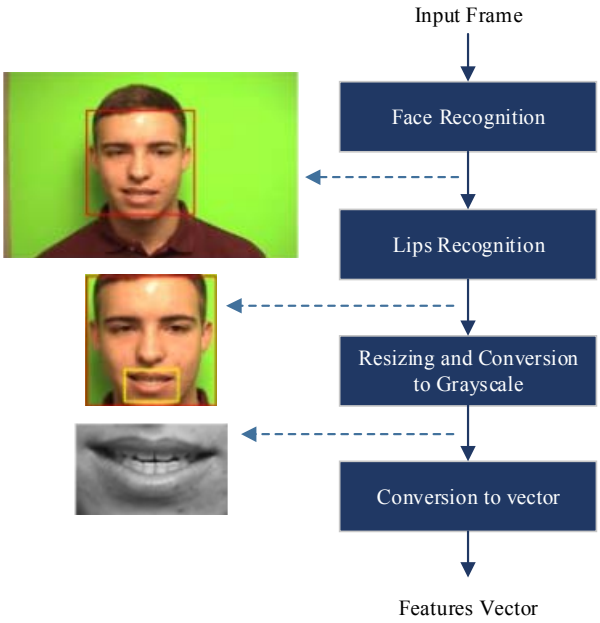

Fig. 1. The procedure of the "raw gray level ROI" feature extraction with an example frame.

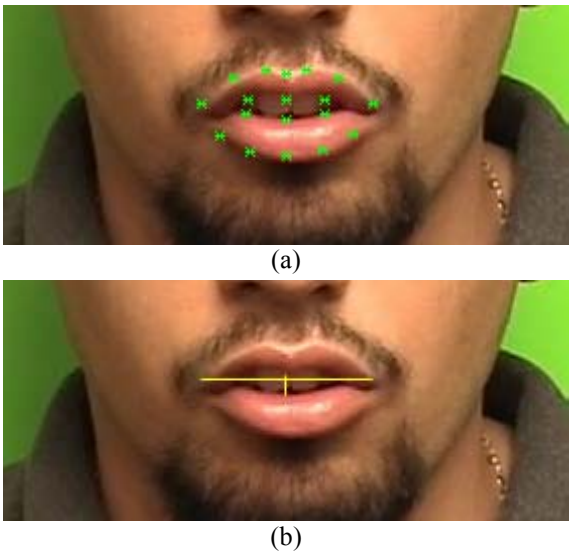

Fig. 2. a. 18 landmarks of the lips which are used as feature vector. **b.** width and height of the lips extracted from the corresponding landmarks and concatenated to the feature vector.

The generated feature vector obtained via this approach is rich enough to reconstruct its high dimensional input image, with a low measured Root Mean Square Error (RMSE) criterion.

Architecture of the proposed DANN is shown in Fig. 3. The activation function of the output layer is linear and of all the other layers is the log sigmoid function. The DANN has two major parts: "Analysis" and "Synthesis". The analysis part consists of three layers which lead to the bottleneck of the network. This part analyses the input to extract its nonlinear principal components that include the abstract information of the image [10]. The synthesis part also consists of three layers with the bottleneck layer as its input. This part reconstructs the analyzed input image from its 10-dimensional extracted nonlinear principal components. The DANN is firstly pre-trained through the layer-by-layer procedure [11]. Finally, a fine tuning process is applied over the pre-trained network to reduce the remained RMSE in the reconstructed image against the original one. The well-known error Back Propagation (BP) algorithm is employed to teach 3066 lips ROIs of 28 different speakers (15 males and 13 females) to the DANN, by employing the selected unsupervised learning method. The proposed DANN is simulated, trained and tested using the MATLAB toolbox.

B. Speech recognition system

The HMM is employed as the visual speech classifier, due to its great performance in modeling temporal processes. Because of the limited number of phonemes that are actually pronounced in the used speech corpus, we have chosen "phonemes" as the units that must be modeled and then recognized. So, we do not consider "visemes" (the visual form of phonemes) as the modeled units, allowing the developed system remain compatible with acoustic features that may be added to the inputs, in our probable future work in the audio visual speech recognition field.

978-1-5090-6454-0/17/$31.00© IEEE

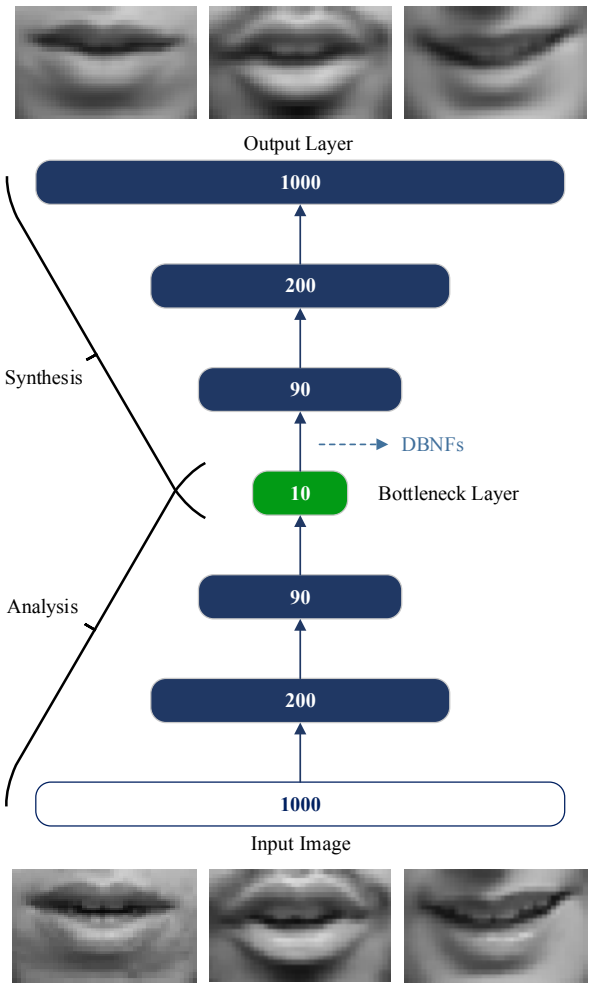

Fig. 3. Architecture of the proposed 6-layer DANN with three examples of experimental lips ROIs.

A DNN with 6 hidden layers is trained to estimate the posterior probabilities of the HMM states [12] along with a conventional Gaussian Mixture Model (GMM) as the baseline system. We also prepare unigram and bigram Language Models (LMs) from the constructed lexicon which affects the weights of lattice trees in the decoding process [13]. All the implementations needed for realization of the mentioned classifier and conduction of the experiments are done using the Kaldi speech recognition toolkit [14].

C. Database

The experiments are done over the CUAVE audio visual database of isolated and connected digits (zero to nine). The database consists of both individual speakers and speaker pairs uttering digits in different situations. The individual section which is of our interest consists of 36 speakers (19 males and 17 females). The speakers are asked to utter the digits in various conditions. First 50 isolated digits are uttered while standing naturally without any movements. Then 30 isolated digits are spoken while moving side-to-side, back-and-forth and tilting the head. In the third and fourth parts, each speaker utters 20 connected digits on both profile views. Finally 60 connected digits including telephone-number-like sequences are spoken while facing the camera again (A half while standing and a half while moving) [5]. Table I, shows the specifications of the database at a glance.

III. Experiments and Results

The individual section of the database is divided into two parts: 86% of the database is devoted to the training and the other 14% to the test set. The third and the fourth part of each speaker data, which are the profile views, are neglected. Although, the moving parts of the videos are included in the experiments, the utterances with incomplete facial parts in the corresponding frames are also neglected due to their hardness of face recognition phase, depicted in Fig. 1.

The digit recognition task is done in the phoneme level including a total of 19 separate phonemes and 1 extra code for silent frames that is not included in the dictionary. A bigram LM is trained over the training set that scores the probabilities of phoneme pairs.

To form the DNN-HMM and the baseline GMM-HMM visual model, firstly for each speaker the feature vectors are normalized to have zero mean and unit variance. A 7-frame context of spliced features of each of the last two introduced kinds (The lips shapes and the DBNFs) are prepared and projected down to 40 dimensions using the Linear Discriminant Analysis (LDA). Additionally, a single Feature-space Maximum Likelihood Linear Regression (FMLLR) transform (estimated per speaker) is utilized to perform Speaker Adaptive Training (SAT). Finally, a set of tri-phonic HMM models are trained over the mentioned resultant feature sets along with GMM model to estimate the emission probabilities of the HMMs in the baseline system.

TABLE I. Specifications of The CUAVE Individual Section Data Samples

General Information		
# of Speakers	36	19 males & 17 females
Information of Each Video		
Video Part	***# of Utterances***	***Details***
1	50	Isolated digits, Without movemets
2	30	Isolated digits, With movements
3	20	Connected digits, Right profile
4	20	Connected digits, Left profile
5	30	Connected digits, Without movements
6	30	Connected digits, With movements

978-1-5090-6454-0/17/$31.00© IEEE

In order to train the DNN, the same feature set as for the HMM train set is used, except that the normalization step is performed globally and the context is prepared on an 11-frame window. A pre-training stage is applied on the Restricted Boltzmann Machines (RBMs) to initialize the weights of the DNN as it is a stack of the pre-trained RBMs. Next, the DNN is fine-tuned to complete the DNN-HMM hybrid model. The model is used as the implemented Visual Model (VM) to generate the VM scores for the lattice generating decoder. Fig. 4 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 are trained over the selected training set.

The described framework is also valid for the experiments conducted using the first introduced feature type (Gray level ROIs) except that, no context is used besides the current frame feature vector, either in the HMM or DNN training phases, due to the huge dimension of input feature vectors.

Table II and Fig. 5, show the final results of the experiments. The numbers show the Phoneme (recognition) Error Rate (PER) percentages measured during the different conducted lip-reading experiments performed on both the test set and a randomly selected subset of the train set which is equal to the test set in volume.

The shape features show improved PERs in comparison to the ROI features in all the four possible conditions (Train and test results of the baseline system and the DNN-HMM hybrid). The results also show significant improvements for the DBNFs in comparison to both the ROI and shape features. For the test set, the DBNFs act relatively 16.2% better than the shape features, using the baseline recognition system, and 14.6% better when the DNN-HMM hybrid is used as the final classifier. Similar condition could be seen for the recognition results obtained over the train set: relatively 16.4% better result is obtained when the baseline system is employed, and 27.5% for the DNN-HMM hybrid application.

Although the baseline system shows better recognition results over the train set, the DNN-HMM hybrid is more successful in recognition of the test set data. This is found in the recognition results obtained over the test set versus the train set, distinctively employing the two implemented systems. This shows that the generalization power of the DNN-HMM hybrid system is superior to the baseline system, which is a considerable conclusion.

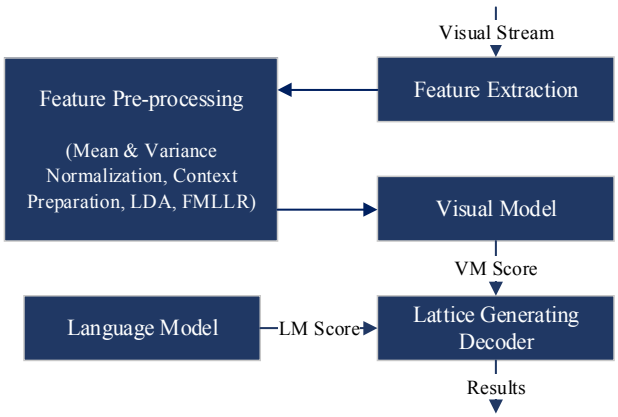

Fig. 4. Overall architecture of the implemented lip-reading system.

TABLE II. RESULTS OF THE EXPERIMENTS IN TERM OF PHONEME ERROR RATES (PER%)

System	Train set			Test set		
	ROI	***Shape***	***DBNF***	***ROI***	***Shape***	***DBNF***
GMM-HMM	21.6%	20.1%	**16.8%**	53.7%	43.7%	36.6%
DNN-HMM	44.6%	34.5%	25%	52.8%	41.1%	**35.1%**

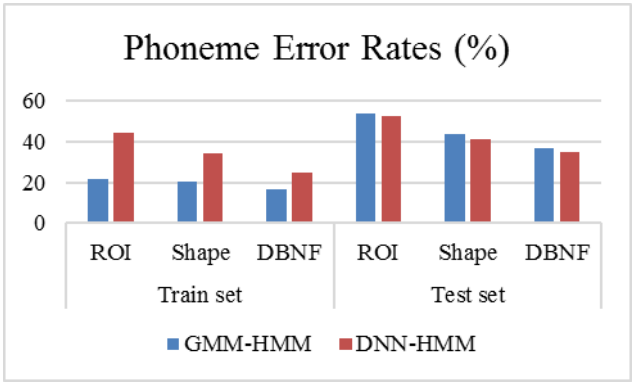

Fig. 5. Results of the experiments in term of phoneme error rates.

IV. CONCLUSIONS

In this paper, we employed three different evaluated feature sets for representing the spoken information found in the video frames recorded from the speakers, inside a speaker independent lip-reading mission. Two HMM-based visual models, including the conventional GMM-HMM and the young DNN-HMM hybrid were implemented to test and compare the introduced features.

After testing the easy-to-access raw gray level ROI of the speakers' lips, the geometric specifications of the lips (the shape features) were employed which showed a lower error rate by 20.4% relative, on average. The DBNFs which benefit from the advantages of both the former feature sets were then employed and showed a relative improvement with an average of 15.4% in comparison to the shape features, over the test data.

ACKNOWLEDGMENT

Special thanks to Mohammad Amin Mahmoudzadeh and Saeed Montazeri Moghadam for their continual guidance in development of this project.

REFERENCES

[1] Foo, S.W., L. Yong, and D. Liang, Recognition of visual speech elements using adaptively boosted hidden Markov models. Circuits and Systems for Video Technology, IEEE Transactions on, 2004. 14(5): p. 693-705.

[2] Lan, Y., et al. Comparing visual features for lipreading. in International Conference on Auditory-Visual Speech Processing 2009. 2009.

[3] Matthews, I., et al. A comparison of active shape model and scale decomposition based features for visual speech recognition. in European Conference on Computer Vision. 1998. Springer.

[4] Luettin, J., N.A. Thacker, and S.W. Beet. Speechreading using shape and intensity information. in Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on. 2002. IEEE.

[5] Patterson, E.K., et al. CUAVE: A new audio-visual database for multimodal human-computer interface research. in Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on. 2002. IEEE.

[6] Stork, D.G., G. Wolff, and E. Levine. Neural network lipreading system for improved speech recognition. in Neural Networks, 1992. IJCNN., International Joint Conference on. 1992. IEEE.

[7] Foo, S.W., Y. Lian, and L. Dong, Recognition of visual speech elements using adaptively boosted hidden Markov models. IEEE Transactions on Circuits and Systems for Video Technology, 2004. 14(5): p. 693-705.

[8] Cootes, T.F., et al., Active shape models-their training and application. Computer vision and image understanding, 1995. 61(1): p. 38-59.

[9] Milborrow, S. and F. Nicolls. Active shape models with SIFT descriptors and MARS. in Computer Vision Theory and Applications (VISAPP), 2014 International Conference on. 2014. IEEE.

[10] Kramer, M.A., Nonlinear principal component analysis using autoassociative neural networks. AIChE journal, 1991. 37(2): p. 233-243.

[11] Seyyedsalehi, S.Z. and S.A. Seyyedsalehi, A fast and efficient pre-training method based on layer-by-layer maximum discrimination for deep neural networks. Neurocomputing, 2015. 168: p. 669-680.

[12] Veselý, K., et al. Sequence-discriminative training of deep neural networks. in Interspeech. 2013.

[13] Mohri, M., F. Pereira, and M. Riley, Speech recognition with weighted finite-state transducers, in Springer Handbook of Speech Processing. 2008, Springer. p. 559-584.

[14] Povey, D., et al. The Kaldi speech recognition toolkit. in IEEE 2011 workshop on automatic speech recognition and understanding. 2011. IEEE Signal Processing Society.

978-1-5090-6454-0/17/$31.00© IEEE