Digital Signal Processing ••• (••••) •••–•••

Contents lists available at ScienceDirect

Digital Signal Processing

www.elsevier.com/locate/dsp

Audio-visual feature fusion via deep neural networks for automatic speech recognition

Mohammad Hasan Rahmani, Farshad Almasganj *, Seyyed Ali Seyyedsalehi

Biomedical Engineering Department, Amirkabir University of Technology (Tehran Polytechnic), Hafez Ave., Tehran, Iran

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Audio-visual speech recognition
Deep autoencoder
Deep neural networks
Feature extraction
Multimodal information processing

ABSTRACT

The brain-like functionality of the artificial neural networks besides their great performance in various areas of scientific applications, make them a reliable tool to be employed in Audio-Visual Speech Recognition (AVSR) systems. The applications of such networks in the AVSR systems extend from the preliminary stage of feature extraction to the higher levels of information combination and speech modeling. In this paper, some carefully designed deep autoencoders are proposed to produce efficient bimodal features from the audio and visual stream inputs. The basic proposed structure is modified in three proceeding steps to make better usage of the presence of the visual information from the speakers' lips Region of Interest (ROI). The performance of the proposed structures is compared to both the unimodal and bimodal baselines in a professional phoneme recognition task, under different noisy audio conditions. This is done by employing a state-of-the-art DNN-HMM hybrid as the speech classifier. In comparison to the MFCC audio-only features, the finally proposed bimodal features cause an average relative reduction of 36.9% for a range of different noisy conditions, and also, a relative reduction of 19.2% for the clean condition in terms of the Phoneme Error Rates (PER).

1. Introduction

The method of utilizing multiple sources of information for better perception of the surroundings is a fundamental question in multimodal information processing which highly affects the performance of such systems [28]. The Audio Visual Speech Recognition (AVSR) task is an instance of the multimodal information processing that exploits two separate information modalities with different characteristics to generate its output. The Automatic Speech Recognition (ASR) process that is basically introduced and developed on only auditory information, shows significant improvements when gets the extra benefits of the associated visual sensory information [12,13,15,23], inspired by the human brain functionality [9].

Employing two various data streams with completely different physical appearances that carry related information, adds to the significance of the AVSR task. To come over the difficulties of implementing such a task, choosing a high-performance model which can truly address the nonlinear correlations between these two different information sources, gets even more important. Moreover, considering varieties in language, accent, race, skin color, face theme, environmental light condition, camera position relative to the face, etc., the AVSR task becomes a more complicated problem needing huge audio-visual databases and consequently powerful computational hardware.

Various methods are used to setup AVSR systems. Many of them are based on the probabilistic models such as the ones derived from the Hidden Markov Model (HMM); many others are constructed via different Neural Network based architectures. There are also lots of ideas connecting the mentioned approaches to implement more powerful AVSR systems.

Neural Networks (NNs) offer many efficient frameworks in all stages of the AVSR systems, including the feature extraction, information fusion, and speech modeling. Different kinds of neural networks are utilized for this purpose [24] especially after the introduction of the novel deep learning methods that have made many utilities for more rapid and successful implementation of the Deep NNs (DNNs) [5,22]. The introduced networks range from the autoencoders (e.g. [14]) and the convolutional neural networks (e.g. [15] and [21]) to the Elman and probabilistic neural networks (e.g. [2]). Moreover, recently, the Long Short-Term Memory (LSTM) networks are successfully employed for the AVSR purposes [17,25].

To model temporal processes, including the human speech, the HMM has shown great performance and is widely used in the ASR

* Corresponding author.
E-mail addresses: mhnrahmani@aut.ac.ir (M.H. Rahmani), almas@aut.ac.ir (F. Almasganj), ssalehi@aut.ac.ir (S.A. Seyyedsalehi).

https://doi.org/10.1016/j.dsp.2018.06.004
1051-2004/

systems along with either the conventional Gaussian Mixture Models (GMMs) [3] or the state-of-the-art DNNs [1,10] as the posterior probabilities estimators needed for the HMM states.

In this paper, the main contribution is to achieve a high-performance audio-visual feature combination structure with respect to its applications in the AVSR task. It is focused on combining the video and audio information through the deep autoencoders applied to both the auditory and visual features. The well-known Mel Frequency Cepstral Coefficients (MFCC) are exploited as the auditory features. As the basic visual features, similar to our previous work [19], the Deep Bottleneck Features (DBNF) [4,6] are extracted via an independent 6-layer deep autoencoder on top of the raw gray-scale lips Region of Interest (ROI). Similar to what Ngiam et al. [14] proposed, the information combination process yields to a shared representation of the two streams and the product is used as the input features to the audio-visual model for further temporal processing. As stated in the previous studies and of course, confirmed in this work, the recognition performance of the ASR systems with video-only inputs (say lip-reading) is inferior to the ones with audio-only inputs [7,14]. Moreover, it is shown that many of the recently developed ASR systems work well when fed by high Signal-to-Noise Ratio (SNR) speech signals (near to clean condition). So, the speakers' videos could be beneficial for noisy environments in which the performance of the audio based ASR systems is dramatically degraded. Consequently, the desired mutual features are produced in a way that the underlying environmental noise in the audio information gets descended in presence of the visual modality. In this regard, the proposed structure in this work benefits from a semi-supervised deep NN by utilizing phoneme labels in the bottle-neck of a bimodal deep autoencoder architecture. The mentioned semi-supervised structure for extracting high-performance bimodal features has never been employed in the previous studies.

In this research, firstly, two unimodal audio and video speech recognition systems are considered to be the baseline systems as well as a simple AVSR system that uses the concatenated audio-visual features as its classifier input. Then, a bimodal deep autoencoder architecture is proposed by which, the nonlinear combination of the features from the audio and video modalities is performed. Next, the proposed architecture and its training procedure are modified in three steps to promote the performance of the extracted bimodal features. All the implemented schemas (including the baselines and the proposed ones) are tested subsequently in noisy conditions; for this purpose, various kinds of auditory noises with different powers are added to the speech signals so that the benefits of the proposed multimodal features in some difficult conditions could be investigated. The experiments are conducted under a professional phoneme recognition scheme which employs the CUAVE database of audio-visual spoken digits [16] and the Kaldi speech recognition toolkit [18].

The rest of the paper is organized as follows: a brief overview of the previous works is presented in section 2. Section 3 introduces the overall implemented system including the blocks used in the experimental setup and the database from which the training and testing sets are selected. Moreover, the feature extraction methods, including the baseline and the proposed ones are discussed in this section. Although the obtained results are presented right after each of the proposed feature combination methods, the complete results and the detailed discussion of them are provided in section 4. Finally, the conclusions are presented in section 5.

2. Related works

Several studies have been done to develop the AVSR systems. They typically propose structures, each improving different parts of the comprehensive AVSR problem. On the other hand, over the last few years, many studies have used the neural networks as a fundamental material in their problem solving. The applications are found in extracting acoustic or visual features, information fusion, denoising distorted information, modeling, classification and final decoding, inside the AVSR scheme. In the following, some of the recently developed related works especially the applications of the DNNs for the AVSR task are reviewed.

Ngiam et al. [14] developed and compared some methods based on the Restricted Boltzmann Machines (RBMs). "Audio RBM", "video-only deep autoencoder" and "bimodal deep autoencoder" were three of the schemes that were tested over CUAVE database. As stated in the paper, the proposed multilayer bimodal network enables the mid-layer features to perceive the nonlinear correlations between the two information branches. In addition to the above schemes, the combinational feature representations of "bimodal + audio RBM" and "video-only deep autoencoder + audio RBM" have been proposed. To evaluate the performance of the five mentioned feature representations, the additional white Gaussian noise was added to the original audio signal at 0 dB SNR. At this SNR, the paper reports its best recognition accuracy from the "bimodal + audio RBM" feature representation (82.2%) while this bimodal representation could not overcome the unimodal "audio RBM" representation in the clean condition (94.4% against 95.8%).

Huang and Kingsbury [7] suggested a fusion on top of mid-level features produced by a bimodal Deep Belief Network (DBN) structure. They also developed decision-fused single-modality DBNs and showed that their feature fusion strategy on top of the mid-level features outperforms their other fusion strategies, including the "DBN decision fusion" and a "baseline fusion implemented using the Multi-Stream HMM (MSHMM)/GMM". The evaluations were performed on a continuously spoken digit recognition task. They reported a relative reduction of 21% in word error rate against the mentioned baseline.

Noda et al. [15] used two applications of the DNNs in their AVSR work. For the acoustic data, they developed a deep denoising autoencoder applied to the input MFCC feature vectors. The autoencoder was trained to filter out the distorted inputs and to predict the clean MFCC features. On the other hand, a Convolutional Neural Network (CNN) was trained to produce phoneme labels from the motion of the speakers' lips images. The output of the CNN was used as the extracted visual features. Finally, the MSHMM was utilized to perform the information fusion and classification. They reported that in an isolated word recognition task, the attained denoised audio features give significant noise robustness and the produced visual features act better than the conventional image-based features such as the Principle Component Analysis (PCA).

Tian et al. [25] proposed an end-to-end deep model called Auxiliary Multimodal LSTM (am-LSTM) network to overcome the weaknesses of the other DNN systems. The am-LSTM consists of two LSTMs (one for audio and the other one for video) and some other components. In the fusion level, the data enters to a projection layer after the two LSTMs and finally, a multi-layer perceptron is used for classification. They reported better results against the ones obtained via the "multimodal deep autoencoder", "multimodal deep belief network" and "recurrent temporal multimodal RBM" approaches.

3. Materials and methods

3.1. Overall framework

The feature extraction techniques, the feature preprocessing stage, the acoustic model, the language model and the lattice generating decoder, are the main blocks of the employed ASR framework. This work has focused on the feature combination

Please cite this article in press as: M.H. Rahmani et al., Audio-visual feature fusion via deep neural networks for automatic speech recognition, Digit. Signal Process. (2018), https://doi.org/10.1016/j.dsp.2018.06.004