Robust Audio-Visual Speech Recognition System based on Gabor Features and Dynamic Stream Weight Adaption

Ali Saudi
Digital Media Engineering and Technology Department
Faculty of Media Engineering and Technology
German University in Cairo, Egypt
Email: ali.saudi@guc.edu.eg

Mahmoud Khalil and Hazem Abbas
Computer and Systems Engineering Department
Faculty of Engineering
Ain Shams University, Cairo, Egypt
Email: {mahmoud.khalil, hazem.abbas}@eng.asu.edu.eg

***Abstract*—This paper aims to enhance the performance of audio-visual speech recognition (AVSR) systems by introducing contributions in both the front-end and back-end system stages. Identifying a reliable feature is a crucial step towards enhancing the front-end stage of both audio-module and visual-module. A two-dimensional Gabor filter with different scales and directions is utilized to generate a set of noise robust spectro-temporal audio and visual features. The performance achieved from the Gabor audio features (GAFs) and Gabor visual features (GVFs) is compared to the performance of the traditional audio features such as MFCC, PLP, RASTA-PLP and visual features such as DCT2. The experimental results demonstrate that a system utilizes Gabor features in the front-end has a much better performance, especially at low SNR levels. To enhance the back-end stage, a framework based on synchronous multi-stream hidden Markov model is proposed to solve the dynamic stream weight estimation problem. To demonstrate the effect of dynamic weighting on enhancing the AVSR performance, we empirically compare between late integration (LI) and early integration (EI) strategies, especially in a low-SNR scenario. The experimental results show that the AVSR-LI system achieves superior performance for all SNR levels compared to AVSR-EI system.**

Keywords*—*Audio-Visual Speech Recognition, Synchronous Multi-Stream Hidden Markov Model, Stream Weight.

I. Introduction

The performance of most speech recognition systems decrease dramatically when distortion or noise affect the audio signal. An approach to resolve this issue is to add another modality information, usually, the visual speech information coming from the lip movements, to complement the audio information. The fusion of visual speech information with audio information will create an audio-visual speech recognition (AVSR) system. The fusion of the audio and visual information can be early applied at the feature level, it's called feature fusion scheme or early integration (EI). It can also be fused at a late level in the decision fusion scheme or late integration (LI). This bi-modal system should provide superior performance compared to both uni-modal systems, the audio-only (ASR) and visual-only (VSR) speech recognition systems. Moreover, it improves the speech recognition system's robustness, especially, under noisy conditions.

However, there are two challenges in most AVSR systems. The first is the process of selecting and extracting the relevant information from each modality. For this purpose, different schemes to extract different audio and visual feature are proposed. In this paper we introduce a set of robust features that encode the spectro-temporal changes across frequencies and over time. This by utilizing the Gabor filter in both audio and visual front-ends.

The usage of 2D-Gabor filter as a feature extractor is a vital component in numerous applications such as speech recognition [1], object recognition [2], image recognition [3], [4], and face recognition [5]. The aim of this process is to find out experimentally the impact of the filterbank type and its effect on the robustness of speech recognition systems in noisy conditions. This can be done by searching for the optimum employment of the two-dimensional (2D) Gabor functions with different scales and directions to analyze both audio and visual speech information.

The performance obtained from utilizing both Gabor audio features (GAFs) and Gabor visual features (GVFs) is compared to the performance of utilizing the conventional features such as Mel-Frequency Cepstral Coefficients (MFCC), Perceptual Linear Prediction (PLP), Relative Spectral Transform (RASTA-PLP) audio features and DCT2 visual features. To quantify the robustness of the proposed AVSR system against extrinsic variation resulted from different noise levels, a series of recognition experiments with different feature extraction algorithms, dimensionality values, and SNR levels were executed.

The second is the process of selecting and adapting an appropriate fusion model, such that the multi-modal fused model introduce a higher performance outperforming both of uni-modal models. In this paper, two fused multi-modal models, namely EI and LI, are considered and compared. In the EI model, a single-stream hidden Markov model (HMM) is trained on vectors of concatenated audio-visual features. In the LI model, a multi-stream HMM (MSHMM) is used. The MSHMM model incorporates weighting of the individual stream likelihoods by the stream weights. The weights are intended to represent the stream confidence.

In this work, we propose a dynamic stream weight adaption method based on the estimated confidence of each stream. A measure of stream confidence is derived from the speech signal itself using six basic features. The confidence measures implicitly estimate the SNR in the audio signal, and

on consequence, the audio stream weight can be related to the stream SNR. The audio with a low SNR would be receiving lower weight, putting more emphasis on the video stream information which is not corrupted by the acoustic noise.

The multi-layer perceptron (MLP) classifier is proposed in order to map the signal-based confidence measures to dynamic stream weights. Finally, we experimentally show the preference of LI approach compared to the EI approach especially in a situation when one of the modalities is corrupted by noise.

The rest of this paper is organized as follows. Section II concentrates on the front-end stage the AVSR system. Section III explains how the audio-visual fusion takes place in speech recognition systems. Section IV explains the stream weight estimation using the proposed dynamic weighting scheme. Section V discusses our audio-visual database and reports experimental results. Finally, Section VI presents conclusions and a brief discussion of the possible directions for future work.

II. The Front-End Stage

A. The Audio Front-End

Most speech recognition systems analyze the short-term spectrum of speech, typically calculated from 20 ms frames. The spectral analysis is performed on these frames as a feature input. The popular techniques of such spectral analysis are MFCCs, PLPs, and RASTA-PLP. These parameters are usually complemented with their first and second order derivatives to encode the temporal changes of the signal.

Physiologically in the auditory system, the primary auditory cortex contains a set of neurons, which are sensitive to different patterns in the spectro-temporal representation of the signal [6]. The estimation of this representation that stimuli the neuron's response is called the spectro-temporal receptive field (STRF). 2D Gabor filters have been used to model STRFs, which motivated the usage of GAFs as a front-end for Audio-only speech recognition (ASR) system. GAFs are computed by processing the spectro-temporal representation of the audio signal by a set of 2D modulation filters. Filtering is carried out by computing the 2D convolution of the filter and a log Mel spectrogram. An implementation of the GAFs calculation steps is available online [7]. Empirically, no significant change was observed in the ASR accuracy when reducing the feature space to be the first 90 GAFs.

B. The Visual Front-End

This stage consists of two main steps, the pre-processing step, and the feature extraction step [8]. In the pre-processing step, the Viola-Jones algorithm [9] is applied on 2D images twice. First, to detect the subject's face and then to extract the mouth region of interest (ROI). The mouth rectangle is rescaled to be 128×128 pixels, then it is transformed to a gray-scale format. Finally, the image-transformed-based features using the conventional DCT2 and the Gabor filter are extracted from the mouth ROI of the subject.

The 30 DCT2 features that have the lowest-frequency were used. It exists in the top left corner of the transformed lip image. These DCT2 features thought to be beneficial in the proposed VSR system as it contains the most relevant lip information.

Gabor filters capable of modeling the the neurooptical response of the primary visual cortex over a range of orientations and spatial frequencies of the images [5], [10].

When 40 Gabor filters are applied on mouth images, the total dimension of the visual feature vector will be 655,360 features. After removing the redundant information from the feature vector, the feature vector size becomes 160. Only 30 GVFs were retained, as there is no noticeable improvement in terms of the accuracy of Visual-only speech recognition (VSR) after this value. The linear interpolation scheme is used to up-sample the visual features frame rate to reach the audio features frame rate.

III. Audio-Visual Features Integration

There are many integration approaches to integrate the multi-modal features, such as the audio-visual features in this work. The easiest integration approach is the EI approach where the audio and the visual modalities are stacked up to form a single audio-visual feature vector. The other fusion method is called LI or decision fusion. In this fusion method, to calculate the final classification result, the class conditional log-likelihoods outputs of the two classifiers are merged using suitable stream weight. The audio stream weight λ_a, and the visual stream weight λ_v are constrained to the following [8]:

$$0 \leq \lambda_a, \lambda_v \leq 1, \quad and \quad \lambda_a + \lambda_v = 1 \tag{1}$$

In the current, two HMM structures were utilized in the experiments; HMM and MSHMM structures. The HMM structure is utilized in the early integration experiments. The MSHMM structure is utilized in the late integration experiments. The current work experiments are performed on the EI and LI approaches.

IV. Stream Weight Estimation

The first phase in the stream weight estimation process is to obtain the most important features of the stream confidence measures. In the current work, the audio stream weight depends on the estimated SNR in the audio signal. Particularly, the audio with a low estimated SNR value would be receiving lower audio stream weight, putting more emphasis on the video stream information which is not corrupted by the acoustic noise. To implicitly estimate the SNR found in the audio signal, a six basic frequency-domain and time-domain features were extracted from the acoustic frame and utilized as audio stream confidence measures. These features are the Spectral Centroid (SC), the Spectral Roll-Off (SR), the Spectral Flux (SF), the Short Time Energy (STE), the Energy Entropy (EE), and the Zero-Crossing Rate (ZCR).

In the second phase, we want to find the optimal stream weight for a certain SNR level. The optimal stream weight can be found by training the AVSR system with a range of possible weight values and then selecting the stream weight that achieves the best AVSR recognition rate. This tedious

and computationally overwhelming process is performed only at the training stage of the AVSR system. Afterwards, the stream weights are instantaneously produced during the testing stage.

In the third phase, we want to map the stream confidence measures to stream weight. The MLP is utilized for this purpose. The proposed stream confidence measures as well as its corresponding optimal stream weight are introduced to train the MLP. The MLP architecture which produces the best performance is selected. This architecture was obtained empirically. The proposed MLP architecture is composed of an input layer, two hidden layers, and one output layer. There are 6 neurons in the input layer, 100, and 10 neurons for the first and second hidden layers, respectively. The output layer has only one neuron for the optimal stream weight.

Last but not least, in the testing stage, the test samples confidence measures are introduced to the MLP in order to estimate the stream exponents. The proposed MLP attained a recognition accuracy of 99.2% .

V. Experimental Results

A. Data Corpus

In the this work, All experiments have carried out on the Clemson University Audio-Visual (CUAVE) corpus [11]. The CUAVE corpus contains 36 speakers of different sex (19 male and 17 female) uttering both connected and isolated digits 0-9 and repeated five times. The videos are recorded in front of a green background and there are both frontal and side views. The experiments were conducted on the frontal connected-digits part of the corpus. The noisy recordings have been artificially created to simulate the audio corruption. This by adding white Gaussian noise signals to the clean audio recordings at seven SNR levels between -10dB and 20dB. The white noise signals stem from the NOISEX-92 collection. The training set consists of 1200 samples and the testing set consists of 240 samples. The HMMs are left-to-right linear models with varying number of states, from 3 to 15 states. The state conditional probabilities are one-component diagonal covariance Gaussian mixtures.

B. Audio-only and Visual-only Recognition

Fig. 1 shows the best minimum Word Error Rate (WER) results of ASR models using MFCC, PLP, RASTA-PLP and GAFs at various SNR levels. It can be seen that GAFs complemented with the first derivative attained a 1.11% WER which significantly outperforms the conventional audio features, especially for higher noisy conditions. Also, it is notable that the WER of the ASR is increasing at a high rate when the SNR level descends. The best ASR performance is attained when the GAFs complemented with the first derivative are utilized as the audio features. Therefore, it is used for following AVSR experiments.

Regarding the VSR system, we compare the performance obtained after utilizing the GVFs, DCT2, and a hybrid of both features. The results demonstrate that the utilization of GVFs features attained a remarkable improvement in the VSR system word recognition performance.

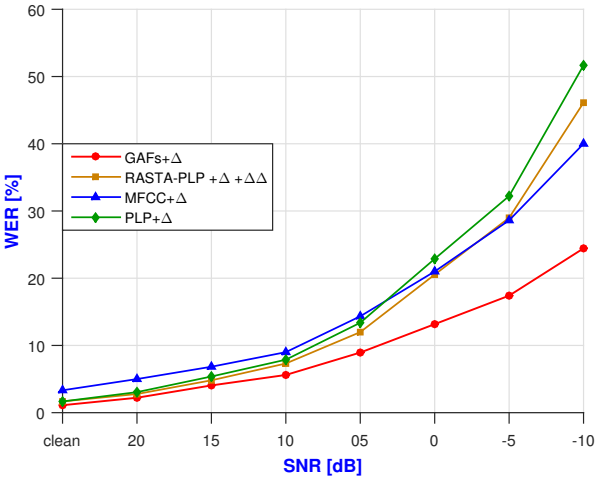

Fig. 1. The resulting word error rates for the ASR after using MFCC, PLP, RASTA-PLP, and GAFs at different SNR levels.

The utilization of GVFs in the VSR model achieved 30.77% WER compared to 46.15% and 38.46% for the DCT2 and hybrid features, respectively. The results demonstrate that the performance of the VSR is improved when utilizing the proposed GVFs feature in the visual front-end compared to the conventional visual features such as DCT2.

C. Audio-Visual with Early Integration Recognition

In the EI experiments, the proposed GAFs complemented with its first derivative were utilized as audio features as well as the DCT2 and GVFs, with their first derivatives as visual features. The single-stream HMM is trained and tested on vectors of concatenated audio-visual features. Fig. 2 shows the best minimum WER results of the AVSR-EI system experiments.

As illustrated in Fig. 2, the utilization of a concatenated feature vector consisted of (GAFs + Δ) + (GVFs + Δ) attained 1.11% of WER in clean environment. In addition, the WER reached to 17.07% and 23.38% at -5dB and -10dB SNR levels, respectively. This attains a 0.71% and a 1.07% absolute improvement difference compared to a single feature vector consisted of (GAFs) + (GVFs) in the same SNR levels conditions.

D. Audio-Visual with Late Integration Recognition

In the LI experiments, the MSHMM setup is utilized to incorporate weighting of the individual stream likelihoods by the stream weights. Fig. 3 shows the best minimum WER results of the AVSR-LI system experiments.

It demonstrates the performance gain of using (GAFs + Δ) as audio features and (GVFs + Δ) as visual features in the AVSR-LI system. The results show that the utilization of (GAFs + Δ) and (GVFs + Δ) attained 1.11% of WER in clean environment which is the highest recognition rate compared to other methods. In addition, the WER reached to 14.29% and 20% at -5dB and -10dB SNR levels, respectively. This attains a 0.89% and a 1.11% absolute

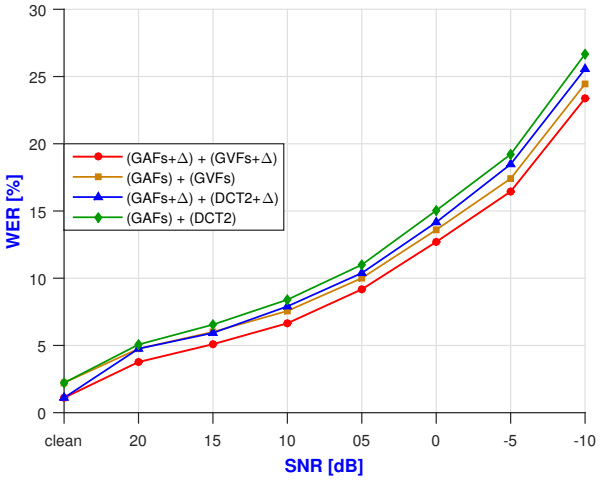

Fig. 2. The resulting word error rates for the AVSR-EI after GAFs as audio features with GVFs, and DCT2 as visual features at different SNR levels.

improvement difference difference in compared to to the performance obtained after of utilizing GAFs as audio features and GVFs as visual features.

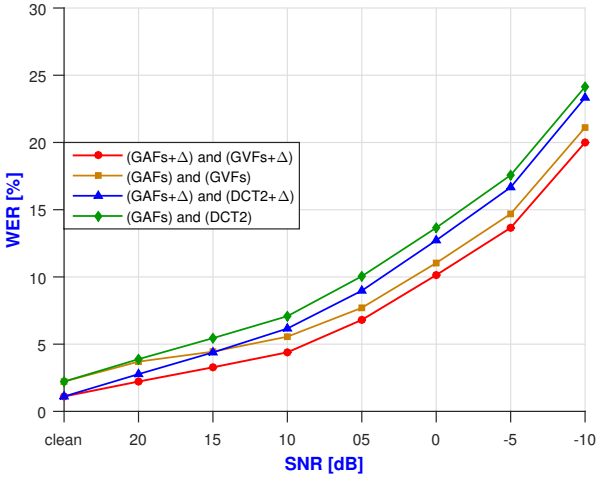

Fig. 3. The resulting word error rates for the AVSR-LI after GAFs as audio features with GVFs, and DCT2 as visual features at different SNR levels.

As illustrated in Fig. 2 and Fig. 3 graphs, the visual features are very useful in improving the speech recognition performance, especially at low SNR levels, compared to the performance obtained from either the ASR or VSR.

The graphs of Fig. 4 demonstrate the performance gain of the proposed weighting approach in improving the multimodal AVSR-LI system compared to the unimodal ASR and VSR systems as well as the multimodal AVSR-EI system. The results show that the proposed weighting approach achieved a remarkable improvement in the AVSR system recognition performance , outperforming both the ASR and AVSR-EI systems by a large difference by reducing the WER from 15.55% to 13.33% with approximately 14.27% relative improvement and from 14.49% and 13.33% with approximately 8% relative improvement, respectively. The results confirm that multimodal AVSR-LI system outperformed all other systems under both clean and noisy conditions.

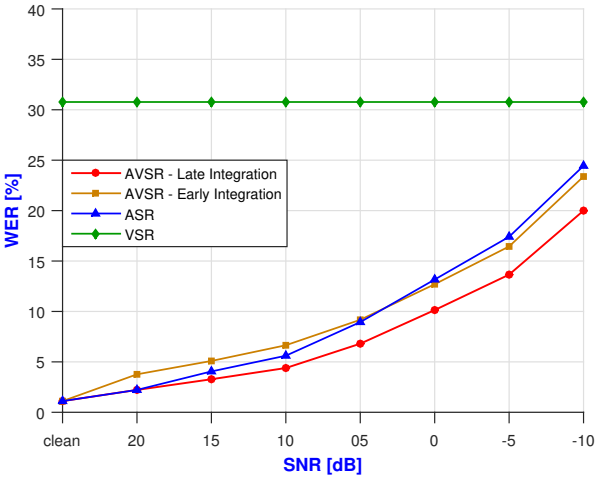

Fig. 4. Performance comparison of ASR, VSR, AVSR-EI, and AVSR-LI systems at different SNR levels.

VI. Conclusions

In this work, the problem of speech feature extraction in noisy environments is investigated. The spectro-temporal information resulted from utilizing Gabor filters used to improve the performance of the ASR and VSR systems. The results demonstrated a remarkable improvement in the performance when utilizing the GAFs and GVFs compared to the conventional audio and visual feature. In addition, the problem of dynamic stream weighting is investigated, and the advantages of LI strategy over EI strategy were addressed. The results confirm that proposed AVSR-LI model that utilizes the dynamic weighting scheme outperform all other models under both clean and noisy conditions. In the future, we will apply the proposed Gabor features on deep learning architectures to solve large vocabulary tasks.

References

[1] J. Schröder, S. Goetze, and J. Anemüller, "Spectro-temporal gabor filterbank features for acoustic event detection," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 23, no. 12, pp. 2198–2208, 2015.

[2] H. Yao, L. Chuyi, H. Dan, and Y. Weiyu, "Gabor feature based convolutional neural network for object recognition in natural scene," in *Information Science and Control Engineering (ICISCE), 2016 3rd International Conference on*. IEEE, 2016, pp. 386–390.

[3] S. S. Sarwar, P. Panda, and K. Roy, "Gabor filter assisted energy efficient fast learning convolutional neural networks," *arXiv preprint arXiv:1705.04748*, 2017.

[4] S. Luan, C. Chen, B. Zhang, J. Han, and J. Liu, "Gabor convolutional networks," *IEEE Transactions on Image Processing*, 2018.

[5] S. Meshgini, A. Aghagolzadeh, and H. Seyedarabi, "Face recognition using gabor-based direct linear discriminant analysis and support vector machine," *Computers & Electrical Engineering*, vol. 39, no. 3, pp. 727–745, 2013.

[6] N. Mesgarani, S. David, and S. Shamma, "Representation of phonemes in primary auditory cortex: how the brain analyzes speech," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 4. IEEE, 2007, pp. IV–765.

[7] M. Schädler, "Gabor filter bank (GBFB) feature extraction reference implementation in matlab," 2011.

[8] D. Stewart, R. Seymour, A. Pass, and J. Ming, "Robust audio-visual speech recognition under noisy audio-video conditions," *IEEE transactions on cybernetics*, vol. 44, no. 2, pp. 175–184, 2014.

[9] P. Viola and M. J. Jones, "Robust real-time face detection," *International journal of computer vision*, vol. 57, no. 2, pp. 137–154, 2004.

[10] L. Shen, L. Bai, and M. Fairhurst, "Gabor wavelets and general discriminant analysis for face identification and verification," *Image and Vision Computing*, vol. 25, no. 5, pp. 553–563, 2007.

[11] E. K. Patterson, S. Gurbuz, Z. Tufekci, and J. N. Gowdy, "Cuave: A new audio-visual database for multimodal human-computer interface research," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Orlando, Florida, USA*, vol. 2. IEEE, 2002, pp. II–2017–II–2020.