Combination of Standard and Complementary Models for Audio-Visual Speech Recognition

Gonzalo D. Sad, Lucas D. Terissi, and Juan C. Gómez

Lab. for System Dynamics and Signal Processing, Universidad Nacional de Rosario
CIFASIS-CONICET, Rosario, Argentina
E-mail: {sad, terissi, gomez}@cifasis-conicet.gov.ar

Abstract. In this work, new multi-classifier schemes for isolated word speech recognition based on the combination of standard Hidden Markov Models (HMMs) and Complementary Gaussian Mixture Models (CGMMs) are proposed. Typically, in speech recognition systems, each word or phoneme in the vocabulary is represented by a model trained with samples of each particular class. The recognition is then performed by computing which model best represents the input word/phoneme to be classified. In this paper, a novel classification strategy based on complementary class models is presented. A complementary model to a particular class j refers to a model that is trained with instances of all the considered classes, excepting the ones associated to that class j. The classification schemes proposed in this paper are evaluated over two audio-visual speech databases, considering acoustic noisy conditions. Experimental results show that improvements in the recognition rates through a wide range of signal to noise ratios (SNRs) are achieved with the proposed classification methodologies.

Keywords: Speech Recognition, Audio-Visual Information Fusion, Decision Level Fusion, Complementary Models.

1 Introduction

Communication among humans is inherently a multimodal process, in the sense that to transmit an idea not only is important the acoustic speech signal but also the visual information during speech, such as mouth movements, facial and body gestures, etc. [4]. This fact has made Audio-Visual Speech Recognition (AVSR) systems a fundamental component in Human Computer Interfaces (HCIs). AVSR systems make use of both acoustic and visual information during speech to perform the recognition task. Several techniques have been proposed in the literature to combine (or fuse) the audio and the visual information [7]. According to the way the information is combined, the techniques can be classified in: those based on Feature Level Fusion [2][6], those based on Classifier Level Fusion [5], and those based on Decision Level Fusion [3][6].

This paper describes new multi-classifier schemes for isolated word speech recognition based on the combination of standard Hidden Markov Models (HMMs) and Complementary Gaussian Mixture Models (CGMMs). In contrast to the case of standard HMMs, where each class is represented with a model trained with instances of the corresponding class, CGMMs are trained with samples of all the remaining classes. For instance, let consider a vocabulary composed by four classes, a, b, c and d, the complementary model to class a is trained with samples of classes b, c and d. In particular, two classification schemes are proposed in this paper to handle data represented by single and multiple feature vectors, respectively. For the case of single feature vectors, a cascade classification scheme using HMMs and CGMMs is proposed. On the other hand, for the case when data is represented by multiple feature vectors, a decision level fusion strategy is proposed. Two audio-visual databases are employed to test the proposed recognition schemes. To evaluate the robustness of the proposed methods, experiments under noisy acoustic channel are performed. The experimental results show that significant improvements in the recognition task are achieved by the proposed classification methods.

The rest of the paper is organized as follows. The proposed classification schemes are described in sections 2, 3 and 4. The description of the audio-visual databases is presented in section 5. In section 6, the experimental results of the proposed systems are shown. Finally, some concluding remarks and perspectives for future work are included in section 7.

2 Complementary models

Typically, the classifiers used in an AVSR systems are implemented using HMMs. In the training stage, an HMM is trained for each particular word in the vocabulary, using several instances of the word to be represented. Then in the recognition stage, given an observation sequence associated with the input word to be recognized, the probability of each HMM is computed and the recognized word corresponds to the HMM who gives the maximum probability. This decision rule is expressed in equation (1), where i is the recognized class, λ_j is the j-th model and O is the observation sequence.

$$i = \underset{j}{\mathbf{argmax}} \;\; P\left(O|\lambda_j\right) , \tag{1}$$

In this paper, another way of using the data in the training stage to produce a set of probabilistic models, namely Complementary Gaussian Mixture Models, is introduced. For each particular word in the vocabulary, a CGMM is trained using all the instances of the words in the vocabulary excepts the corresponding to the one being represented. Then, given an observation sequence associated to the i-word in the vocabulary, if the probability of each CGMM is computed, it would be reasonable to expect that the minimum value would correspond to the i-model. This is because the data used in the training of the i-model doesn't include instances of the i-word, whereas the rest of the models do so. Based on this, in the recognition stage the likelihood score of each CGMM is computed and the recognized word corresponds to the classifier giving the minimum score.

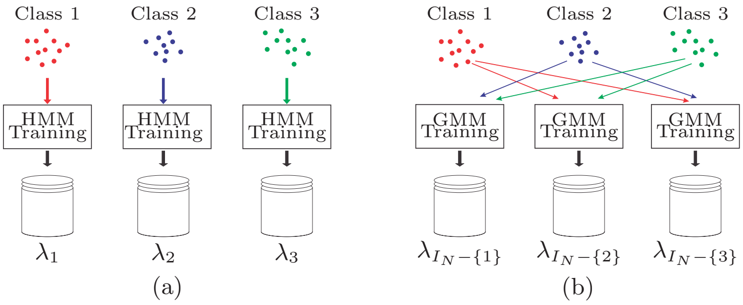

Fig. 1. Training procedure (N = 3). (a) Standard models. (b) Complementary models.

This decision rule can be expressed as follows

$$i = \mathop{\mathbf{argmin}}_{j} \; P\left(O|\lambda_{I_N-\{j\}}\right) \qquad ; \qquad I_N = \{1, 2, 3, \ldots, N\}, \tag{2}$$

where i is the recognized class, N is the vocabulary size, $\lambda_{I_N-\{j\}}$ is the model which has been trained with all the classes in I_N except the j class, which will hereafter be referred to as complementary model. Figure 1 schematically depicts the training procedure for the classifiers based on standard models and complementary models, for the case of a vocabulary with N = 3.

3 Cascade classifiers combination

A combination of traditional and complementary models, using a cascade configuration, is proposed to improve recognition rates. To this end, the recognition is carried out in two stages. First, the M most likely classes are pre-selected using the likelihood scores provided by the λ models. At this point, the possible solutions are reduced to these M classes. Then, the $\lambda_{I_M-\{j\}}$ complementary models of these M selected classes are formed. These models will hereafter be referred to as *M-class complementary models*. Finally, the recognized word corresponds to the $\lambda_{I_M-\{j\}}$ which gives the minimum probability.

Figure 2 schematically depicts the classifier combination strategy proposed in this paper for the case of $M = 3$. Given an observation sequence, associated with the word to be recognized, the λ models are ranked according to their corresponding output probabilities. The $M = 3$ highest ranked models define the classes to be used to form the $\lambda_{I_3-\{j\}}$ complementary models, in this case: $\lambda_{I_3-\{3\}}$, $\lambda_{I_3-\{6\}}$ and $\lambda_{I_3-\{9\}}$. Specifically, $\lambda_{I_3-\{3\}}$ is trained with the training data corresponding to classes 6 and 9, $\lambda_{I_3-\{6\}}$ with the corresponding to classes 3 and 9, and $\lambda_{I_3-\{9\}}$ with the corresponding to classes 3 and 6. Finally, the probability of each complementary classifier is computed and the recognized word corresponds to the classifier who gives the minimum probability.

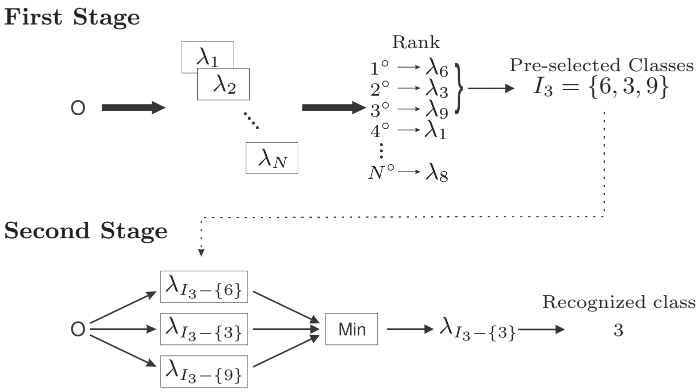

Fig. 2. Example of the proposed classifier combination strategy with $M = 3$.

4 Multimodal fusion strategy

In some cases, the data is represented by multiple features/modalities, each one with its corresponding feature vector. In order to handle this particular situation, a fusion strategy based on a voting scheme is proposed. In Fig. 3, the proposed classification method is schematically depicted for the case of considering data represented by F modalities. It must be noted that now the input of the recognition system are multiple synchronized observation sequences, each one related to each particular modality. For each modality, three decision (class recognitions) are made using classifier based on standard models, complementary models and M-class complementary models, respectively. In particular, in this paper the M-class complementary models are composed considering $M =$ 3. Finally, the individual decisions associated to each particular modalities are combined in a majority vote rule to take the final decision.

5 Audio-visual Databases

The performance of the proposed classification schemes is evaluated over two isolated word audio-visual databases, namely, Carnegie Mellon University (AV-CMU) database (now at Cornell University) [1], and a database compiled by the authors, hereafter referred to as AV-UNR database.

I) *AV-UNR* database: The AV-UNR database consists of videos of 16 speakers, pronouncing a set of ten words (*up*, *down*, *right*, *left*, *forward*, *back*, *stop*, *save*, *open* and *close*) 20 times. The audio features are represented by the first eleven non-DC Mel-Cepstral coefficients, and its associated first and second derivative coefficients. Visual features are represented by three parameters, *viz.*, mouth height, mouth width and area between lips.

II) *AV-CMU* database: The AV-CMU database [1] consists of ten speakers, with each of them saying the digits from 0 to 9 ten times. The audio features are represented by the same parameters as in AV-UNR database. To represent

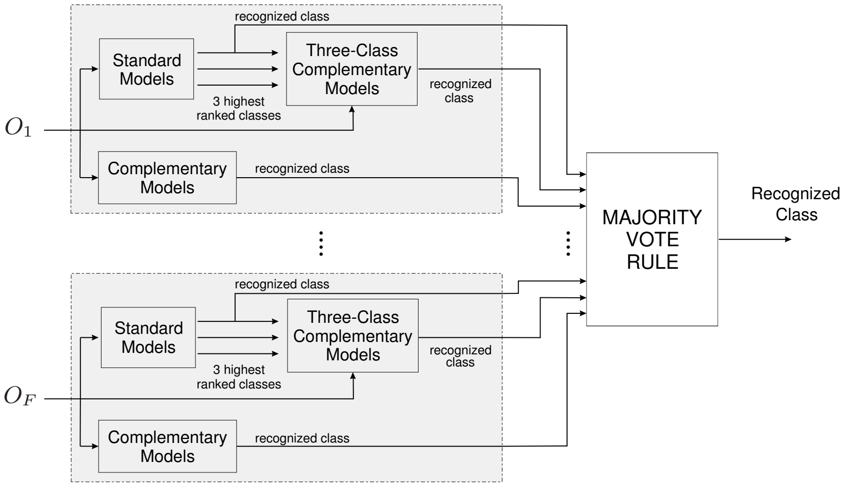

Fig. 3. Fusion strategy for the case of data represented by multiple features/modalities.

the visual information, the weighted least-squares parabolic fitting method proposed in [2] is employed in this paper. Visual features are represented by five parameters, *viz*, the focal parameters of the upper and lower parabolas, mouth's width and height, and the main angle of the bounding rectangle of the mouth.

6 Experimental Results

The proposed classification schemes presented in sections 3 and 4 are tested separately on the databases described in section 5. Experiments with additive Babble noise, added intentionally to the databases, were performed. To obtain statistically significant results, a 5-fold cross-validation is performed over the whole data in each of the databases, to compute the recognition rates. The classifiers based on standard models are implemented using left-to-right HMMs with continuous observations. For the classifiers based on complementary models and three-class complementary models, Gaussian Mixture Models with continuous observations were used. It is important to note that these models were trained using clean data, and the additive noise was injected to the testing sets.

6.1 Audio only data

In Figs. 4(a) and 4(b), the results of the experiments over the two databases for the case of the audio modality, are depicted. Only the medians for each noise level are shown for visual clarity reasons. For the case of the AV-UNR database (Fig. 4(a)), the higher accuracy using standard models was obtained for HMMs with 3 states and 4 Gaussian mixtures, while the best performance of the proposed method based on three-class complementary models was obtained using GMMs with 96 Gaussian mixtures. On the other hand, for the case of

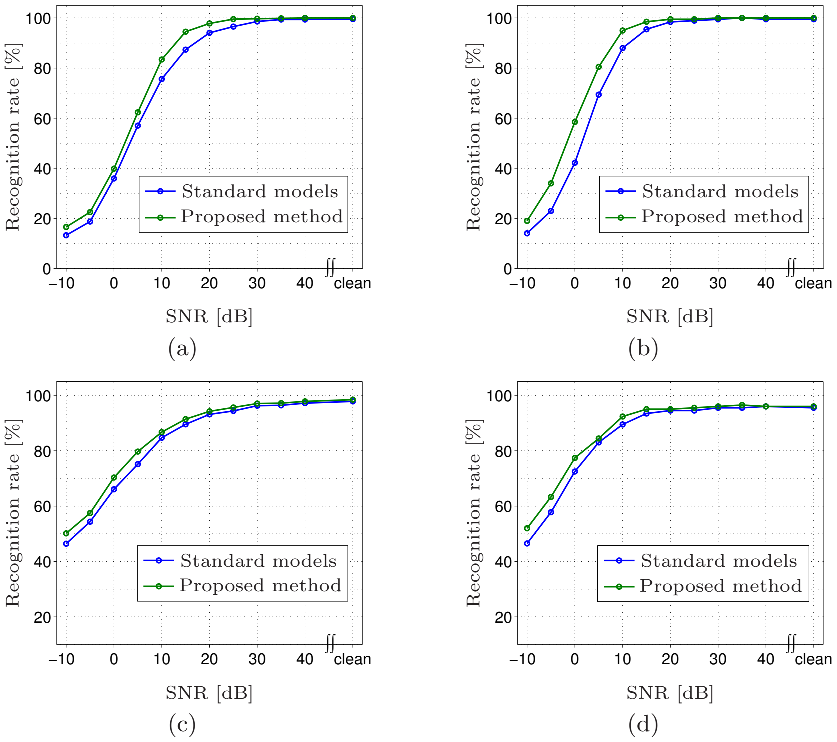

Fig. 4. Recognition rates for the audio ((a) and (b)) and the audio-visual ((c) and (d)) data, using the proposed classifiers. (a) and (c) AV-UNR database. (b) and (d) AV-CMU database.

the AV-CMU database (Fig. 4(b)), the higher accuracy was obtained for HMMs with 3 states and 1 Gaussian mixtures (standard model) and GMMs with 12 Gaussian mixtures (proposed method). As it can be observed, the use of the complementary models improves the recognition over the full range of SNRs.

6.2 Audio-visual data

The classification scheme described in subsection 3 is also evaluated using audio-visual data. Early integration techniques were employed to combine audio and visual information into a single audio-visual vector. The results of the experiments over the two databases for this configuration are depicted in Figs. 4(c) and 4(d). For the case of the AV-UNR database (Fig. 4(c)), the higher accuracy using standard models was obtained for HMMs with 6 states and 4 Gaussian mixtures, while the best performance of the proposed method based on three-class complementary models was obtained using GMMs with 64 Gaussian mixtures. The higher accuracy, for the case of the AV-CMU database (Fig. 4(d)), was obtained for HMMs with 6 states and 1 Gaussian mixtures (standard model) and GMMs with 20 Gaussian mixtures (proposed method). From this figure, it can be observed that the recognition over the full range of SNRs is improved using the complementary models strategy.

6.3 Multi-feature data

When the acoustic channel is corrupted by noise, which is the usual situation in most applications, an improvement can be achieved by fusing audio and visual features. The efficiency of a classifier based on audio-only information deteriorates as the SNR decreases, while the efficiency of a visual-only information classifier remains constant, since it does not depend on the SNR in the acoustic channel. However, the use of only visual information is usually not enough to obtain relatively good recognition rates. It has been shown in several works in the literature [4][7], that the use of audio-visual feature vectors (early integration) improves the recognition rate in the presence of noise in comparison to the audio-only case. Different performances can be achieved depending on the amount of information used for each modality. If the recognition rates of these three classifiers (audio, visual, and audio-visual classifiers) are compared, in most cases occurs that each one performs better than the others in different regions of SNR. Usually, visual classifiers achieve better recognition rates at low SNR, audio classifiers at high SNR, and audio-visual classifiers at middle SNR.

Taking into account the previous analysis, the classification strategies described in 3 and 4 will be used here in combination, aiming at maximizing the efficiency of the recognition system over the different SNRs. The full range of SNRs is split in three regions: from -10dB to -5dB, from 0dB to 15dB and from 20dB to clean. In the lowest region, the visual modality with the classifiers combination technique in section 3 is used. In the middle region, the audio, visual, and audio-visual data are used with the proposed fusion strategy in section 4. In the highest region, the audio modality with the classifiers combination technique in section 3 is used.

The results of these experiments over the two databases, are depicted in Figs. 5(a) (AV-UNR database) and 5(b) (AV-CMU database). In these figures, the recognition rates corresponding to the audio, visual and audio-visual classifiers based on standard models (λ) are also depicted. It is clear that the proposed objective of improving the recognition rates over the full range of SNRs has been accomplished for both databases. In addition, the performance of the proposed system is comparable to that of other methods presented in the literature [7].

7 Conclusions

Novel multi-classifier schemes for isolated word speech recognition based on the combination of standard HMMs and CGMMs were proposed in this paper. In particular, two classification schemes were proposed in this paper to handle data represented by single and multiple feature vectors, respectively. For the case of data represented by single feature vectors a cascade classification scheme using HMMs and CGMMs was presented. In addition, when data is represented

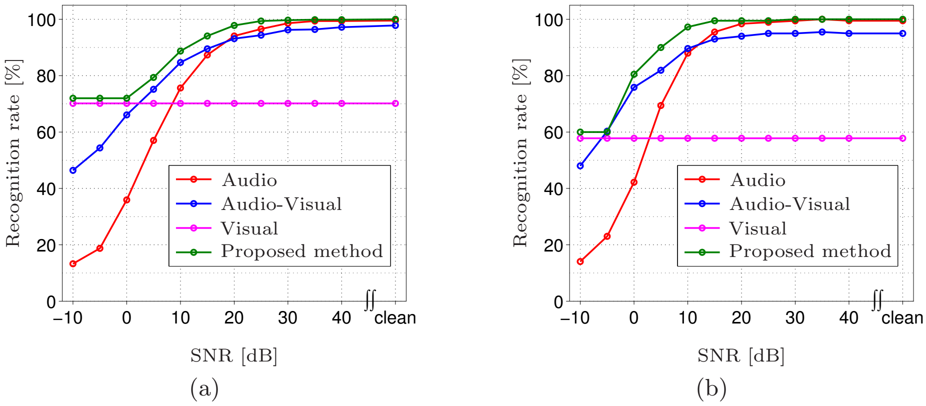

Fig. 5. Recognition rates for the proposed combined strategy. The recognition rates corresponding to the audio, visual and audio-visual classifiers based on standard models (λ) are also depicted. (a) AV-UNR database. (b) AV-CMU database.

by multiple feature vectors, a classification scheme based on a voting strategy which combines scores from individual HMMs and CGMMs was also proposed. These classification schemes were evaluated over two audio-visual speech databases, considering acoustic noisy conditions. Experimental results show that in both cases, the proposed methods lead to improvements in the recognition rates through a wide range of signal-to-noise ratios. Absolute recognition rates could be further improved by including noisy features in the training stage.

References

1. AMP Lab.: Advanced Multimedia Processing Laboratory. Cornell University, Ithaca, NY, `http://chenlab.ece.cornell.edu/projects/AudioVisualSpeechProcessing`, Last visited: April 2015.
2. Borgström, B., Alwan, A.: A low-complexity parabolic lip contour model with speaker normalization for high-level feature extraction in noise-robust audiovisual speech recognition. IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans 38(6), 1273–1280 (2008)
3. Estellers, V., Gurban, M., Thiran, J.: On dynamic stream weighting for audio-visual speech recognition. IEEE Trans. Audio, Speech, Language Process. 20(4), 1145–1157 (2012)
4. Jaimes, A., Sebe, N.: Multimodal human-computer interaction: A survey. Computer Vision and Image Understanding 108(1-2), 116–134 (2007)
5. Papandreou, G., Katsamanis, A., Pitsikalis, V., Maragos, P.: Adaptive multimodal fusion by uncertainty compensation with application to audiovisual speech recognition. IEEE Trans. Audio, Speech, Language Process. 17(3), 423–435 (2009)
6. Sad, G., Terissi, L., Gómez, J.: Isolated word speech recognition improvements based on the fusion of audio, video and audio-video classifiers. In: Proceedings of the XV RPIC. pp. 391–396 (Sept 2013)
7. Shivappa, S., Trivedi, M., Rao, B.: Audiovisual information fusion in human computer interfaces and intelligent environments: A survey. Proceedings of the IEEE 98(10), 1692–1715 (2010)