International Journal of Speech Technology (2022) 25:231–249
https://doi.org/10.1007/s10772-021-09944-7

Complementary models for audio-visual speech classification

Gonzalo D. Sad[1] · **Lucas D. Terissi**[1] · **Juan C. Gómez**[1]

Received: 10 June 2020 / Accepted: 13 November 2021 / Published online: 7 January 2022

Abstract
A novel scheme for disambiguating conflicting classification results in Audio-Visual Speech Recognition applications is proposed in this paper. The classification scheme can be implemented with both generative and discriminative models and can be used with different input modalities, viz. only audio, only visual, and audio visual information. The proposed scheme consists of the cascade connection of a standard classifier, trained with instances of each particular class, followed by a complementary model which is trained with instances of all the remaining classes. The performance of the proposed recognition system is evaluated on three publicly available audio-visual datasets, and using a generative model, namely a Hidden Markov model, and three discriminative techniques, viz. random forests, support vector machines, and adaptive boosting. The experimental results are promising in the sense that for the three datasets, the different models, and the different input modalities, improvements in the recognition rates are achieved in comparison to other methods reported in the literature over the same datasets.

Keywords Speech classification · Audio-visual speech · Complementary models · Classifier combination

1 Introduction

It is of common knowledge that, besides the acoustic signal, the visual information during speech related to facial expressions, hand gesture and body posture contributes significantly to the intelligibility of the message being transmitted, and to the perception of the actual meaning of the message (McGurk & MacDonald, 1976). In addition, as pointed out in a recent survey about the interaction between gesture and speech (Wagner et al., 2014), the parallel use of these modalities gives the listener access to complementary information not present in the acoustic signal by itself. In recent years, the study of human communication has benefited from the increasing number of multimodal corpora available to researchers in this field. Significant research effort has been devoted to the development of Audio Visual Speech Recognition Systems (AVSRS) where the acoustic and visual information (mouth movements, facial gestures, etc.) during speech are taken into account (Jaimes & Sebe, 2007; Katsaggelos et al., 2015; Shivappa et al., 2010).

In recent years, several models have been developed to perform speech recognition by fusioning audio and visual information. One of the first models presented in the literature (and one still widely used) is Hidden Markov Models (HMMs), which results particularly useful because it can handle time series (like speech signals) in a very efficient way (Biswas et al., 2016; Sad et al., 2017; Stewart et al., 2014; Terissi et al., 2015b). Others classical models from the Machine Learning area has been also implemented, like Artificial Neural Networks (ANN) (Potamianos et al., 2003; Savchenko & YaI, 2014), K-Nearest Neighbors (K-NN) (Pao et al., 2009; Shin et al., 2011), matching methods utilizing dynamic programming, Adaptive Boosting classifiers (AdaBoost) (Foo et al., 2004; Schapire & Singer, 1999), Support Vector Machine (SVM) (Benhaim et al., 2014; Vallet et al., 2013), Linear Discriminant Analysis (Potamianos et al., 2001; Zeiler et al., 2016) and Random Forests (RF) (Terissi et al., 2015a, 2018). Most of these models can not handle time series, so a pre-processing step is required to perform some normalization procedure. In recent years, newer and

✉ Gonzalo D. Sad
sad@cifasis-conicet.gov.ar

Lucas D. Terissi
terissi@cifasis-conicet.gov.ar

Juan C. Gómez
gomez@cifasis-conicet.gov.ar

[1] Laboratory for System Dynamics and Signal Processing, Universidad Nacional de Rosario, CIFASIS-CONICET, Rosario, Argentina

more sophisticated models, like Restricted Boltzmann Machines (RBM) (Amer et al., 2014; Hu et al., 2016), Deep Learning (DL) (Guglani & Mishra, 2020; Tao & Busso, 2020) and sparse coding (Ahmadi et al., 2014; Shen et al., 2014), have proved to be very suitable for speech recognition tasks, improving the recognition rates obtained with traditional models like HMMs. This improvement is at the expense of increasing the complexity of the system and the requirement of much more data in the training steps.

All the above mentioned models, can be classified either as generative models or discriminative models. Given an observable variable X and a target variable Y, a generative model is a statistical model of the joint probability distribution P(X,Y), while a discriminative model is a model of the conditional probability of the target Y, given an observation x, symbolically, P(Y|X = x). In speech recognition tasks, the generative models (HMMs, RBM) are formed using one model for each class, i.e., one model for each phoneme or word that compose the dictionary of the problem being analyzed. In the training step, each model is trained using examples of the class to be represented. On the other hand, when a discriminative model is used (RF, SVM, K-NN, ANN) for speech recognition tasks, only one classifier is formed, which internally defines all the classes forming the dictionary.

Many classification and recognition task are usually based on a single model (Noda et al., 2015; Papandreou et al., 2009; Terissi et al., 2018). In any model design, some assumptions and simplifications of the reality are made in order to facilitate the model generation and to mitigate its complexity. Relaxing the model assumptions, may lead to performance improvements, making the model able to represent more complex situations. Also, in order to improve the system's capability to represent complex data, the amount of parameters could be increased. Usually, the system's ability to model complex data is directly related to the number of underlying model parameters. In almost all cases, there exists some training procedure where the underlying model parameters are tuned and the performance of the system is evaluated. Improving the training algorithms, e.g., by using discriminative training, bootstrap methods or boosting methods, is another option to achieve performance improvements (Najkar et al., 2014). However, although all the aforementioned methods can lead to improvements in the performance of the model, they also have side effects or drawbacks such as overtraining, lack of generalization, curse of dimensionality effect, etc. In addition, to implement these methods much more training data are required, which are not always available and in case they are, it is very expensive to have them accurately tagged. A detailed description and analysis of these methods can be found in Breslin (2008).

Another way to achieve improvements in recognition and classification tasks is by resorting to a combination or ensemble of models, instead of trying to enhance a single model (Bilmes & Kirchhoff, 2003; Deng & Li, 2013; Kittler et al., 1998; Krawczyk & Cyganek, 2017; Liu et al., 2004). These improvements can be obtained if the combined models are complementary, in the sense that they deliver different (and complementary) classification results. There are different reasons why it is convenient to employ an ensemble of models instead of just one model (Dietterich, 2000). Since the amount of training data is always limited, a single model will only adjust its estimates to this portion of data within all the universe of true data distribution. So, averaging the output of multiple models may be a better approximation to the true value than the output estimate of a single model. Also, in the training procedure of a single model, the model parameters are tuned until a local maximum is reached, which in most of the cases does not match the global maximum in the space of parameters. Averaging the output estimations of multiple models may be closer to the global maximum than the estimation of a single model. Finally, if the data to be represented is very complex, a single model will fail due to its own capability limitations to model data. An ensemble of models will have more power of data representation than a single model, and it will be able to represent more complex data than a single model.

For several decades, the research on model combination or ensemble of models has been an active area in the Artificial Intelligence and Machine Learning fields (Dietterich, 2000). The different approaches to build complementary models available in the literature, can be divided into two major categories: ad-hoc methods and explicit methods. In the first category, the models of the ensemble are generated altering some variable or algorithm of the model, looking for some diversity among them so that each model delivers different errors (Gales et al., 2006; Hain et al., 2007; Hwang et al., 2007; Stüker et al., 2006). Some examples of this category are: altering the optimization algorithm, using different input features (Koziarski et al., 2017) and bootstrapping the training data. In the second category, the diversity among the models in the ensemble is achieved explicitly in the training procedure (Aggarwal & Dave, 2012; Breslin & Gales, 2009; Hu & Zhao, 2007; Prieto et al., 2015; Puurula & Compernolle, 2010) by training all the models in parallel or in an iteratively fashion. This latest approach usually yields an overfitted model, due to the excessive complexity injected in the training stage. Different works in the literature of Audio-Visual Speech Recognition (AVSR) have proposed ensemble models where the complementary models are generated mostly based in ad-hoc methods.

In this paper, a novel scheme for speech classification tasks based on the combination of traditional and complementary models in a cascade structure, is proposed to improve recognition rates. The diversity among the models is achieved by means of an ad-hoc method, altering the way

the training data is employed in the training procedure. The proposed scheme can be implemented with generative and discriminative models. There are no restrictions about the kind of input features on the proposed method, i.e., it can be employed for lip-reading tasks, where the inputs are visual features, for audio speech recognition, where the inputs are audio features, and also for audio-visual speech recognition, where the inputs are acoustic and visual features previously fusionated (early integration). Given a model or classifier (HMM, RF, SVM), the corresponding complementary model proposed in this paper is generated using the same training procedure as in the original model or classifier, but defining a new set of classes for the training examples, aiming to detect absence of a class. In the complementary models, the i-th class is formed using all the instances in the vocabulary except the corresponding to class i. For instance, let consider a vocabulary composed by three classes, C_1, C_2 and C_3, the new classes are defined as AC_1 which contains all the examples of the classes C_2 and C_3, AC_2 which contains all the examples of the classes C_1 and C_3, and AC_3 which contains all the examples of the classes C_1 and C_2.

Given an example to be classified, the first step in the proposed cascade of classifiers is carried out by making an initial classification by evaluating the traditional model. From the output probabilities of this traditional classifier, only the M most likely classes are pre-selected, and the rest are discarded as possible solutions. If there not exists any pair of conflicting classes in this group of M classes, the recognized class is defined by the highest ranked class from this pre-selected group of classes. Otherwise, if there exists any conflicting classes, a second step is carried out where the complementary models are evaluated and their outputs probabilities are used to determine the recognized class.

The objective of this work is to disambiguate conflicting classes in order to improve the resulting recognition rates. For this purpose, a cascade scheme is proposed considering the novel concept of complementary models.

The performance of the proposed system is evaluated with four different models, viz., one generative model (HMM) and three discriminative models (RF, SVM and AdaBoost). In order to evaluate the performance of the proposed system, different experiments are carried out over two public databases, viz., AVLetters database (Matthews et al., 2002), Carnegie Mellon University (CMU) database (Huang & Chen, 1998), and one database compiled by the authors, hereafter referred to as AV-UNR database. As mentioned before, the proposed method has no restrictions about the kind of input features, so experiments with different input features are carried out, viz., audio features, visual features (lip-reading) and fused audio-visual features (early integration). Since in most of real applications the acoustic channel is affected by noise, the performance of the proposed cascade of classifiers is evaluated injecting different kinds of noise in the acoustic information, viz., Gaussian noise and additive Babble noise (Krishnamurthy & Hansen, 2009). Promising results are achieved with the proposed method, obtaining significant improvements in all the considered scenarios. Also, better results than other methods reported in the literature are achieved in most of the cases over the two public databases.

The remainder of this paper is organized as follows. Section 2 presents a preliminary analysis of the class confusability problem and Sect. 3 introduces and explains the classification based on absent classes. In Sect. 4, a description of the proposed system is given. The databases used for the experiments are described in Sect. 5, and the experimental results and the performance of the proposed strategy is analyzed in Sect. 6. Finally, future works and some concluding remarks are given in Sect. 7 and Sect. 8, respectively.

2 Detection of conflicting classes

In speech recognition tasks, usually the units to classify or recognize are phonemes or words, i.e., each of these units are represented by a class within the classifier. Generally, the errors made by the classifier are not completely random, i.e., it can be observed that usual errors correspond to certain classes or group of classes. This source of error is known as Class confusability or intra-class confusion. For example, in visual speech recognition tasks the majority of errors or confusion is between sequences consisting of the same visemes (a viseme is defined as the smallest visibly distinguishable unit of speech). If we would like to recognize the alphabet letters [A–Z], we would see that the most frequently errors would be between the utterances corresponding to the letters *B* and *P* which are composed of phonemes [B + IY] and [P + IY], respectively. If we map the phonemes to its corresponding visemes, we can see that these words are visually the same and composed of the visemes/p + iy/. Similarly the words *C*, *D* and *T*, which are composed of phonemes [S + IY], [D + IY] and [T + IY], respectively, are composed of the same sequence of visemes, /t + iy/, and are also a major source of errors.

This kind of behavior is easily observable in the confusion matrix obtained from the classification results. If the problem at hand has p classes, the resulting confusion matrix is a matrix with p rows and p columns, with the rows representing the classifier decision and the columns representing the true classes. If the classifier makes no errors, the values of off-diagonal entries will be zeros, but if the problem at hand suffers from Class confusability or intra-class errors, there will be a few off-diagonal elements with high values while the rest will have very low values. So, looking at the position (row and column) of these high value off-diagonal elements we can find the most confusable classes, i.e., the

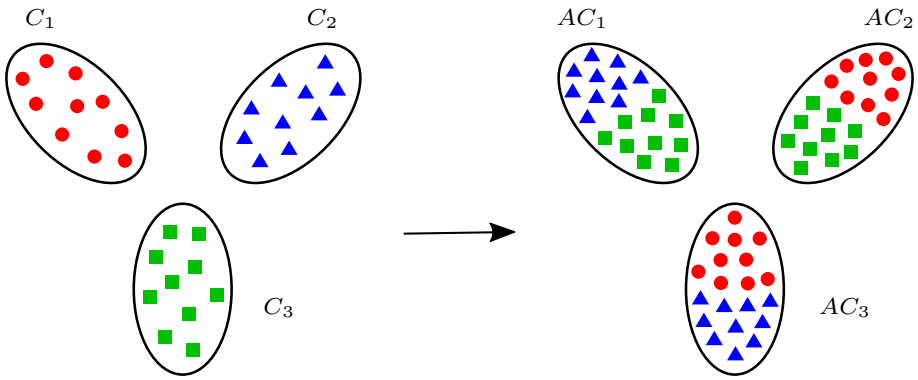


Fig. 1 Procedure to form the new set of classes used for the complementary models. Example for a vocabulary of size N = 3, where $\{C_1, C_2, C_3\}$ are the original classes and $\{AC_1, AC_2, AC_3\}$ are the new ones

classes that are most likely to be confused. In this paper, the P most conflicting classes will be determined by taking the position (row and column) of the P highest values of the off-diagonal elements of the confusion matrix.

3 Complementary models for classification

For speech recognition tasks, a single model is usually used to represent each word or phoneme in a given class when generative models are employed. Each of these models is trained with examples of each particular class. Given the observation sequence to be recognized (O), the models are evaluated and the one given the highest probability of observation is the one that determines the recognized class. When discriminative models are used, a single classifier representing all the classes in the dictionary is trained. Given an observation sequence to be recognized (O), the model is evaluated and the probabilities for each class are obtained. The recognized class corresponds to the one that obtain the highest probability.

An alternative of using the data in the training step to produce a new model is presented in this paper. In contrast to traditional classifiers, the aim of these models is to detect the absence of a class, for this reason the classification technique will be referred to as Complementary model approach. The main idea is to detect the absence of a class by redefining classes, either internally within the model in the case of being discriminative, or individually in the case of being generative. Given a vocabulary composed by N classes, the complementary model is generated by redefining the original N classes. For each class i, a new class is defined, denoted as AC_i, using all the instances in the vocabulary except the corresponding to class i. That is,

$$\begin{cases} AC_i = I_N - \{C_i\} & i = 1, 2, \ldots, N \\ I_N = \{C_1, C_2, \ldots, C_N\} \end{cases} \tag{1}$$

where C_i are the original classes, I_N is the set representing the N original classes, $I_N - \{C_i\}$ is the set containing all the N original classes except the C_i class, and AC_i are the new classes. Figure 1 schematically depicts how the new classes are formed, for the case of a vocabulary of N = 3 classes.

Since the models are trained with the complementary classes AC_i, it would be reasonable to expect that the minimum value of the computed probabilities of an input observation sequence belonging to the i-th class, will correspond to the i-th class. In this way, the recognized word will correspond to the class given the minimum probability, since the model is detecting the absence of the i-th class. The decision rule for the case of discriminative model λ can be expressed as

$$\begin{aligned} i &= \underset{j}{\boldsymbol{argmin}} P\left(O|\lambda_{AC}, AC_j\right) \\ AC_j &= I_N - \{C_j\} \\ I_N &= \{C_1, C_2, \ldots, C_N\}, \end{aligned} \tag{2}$$

where i is the recognized class, λ_{AC} is the classifier which has been trained with the new set of classes (AC_j), which will hereafter be referred to as *complementary discriminative model*, and $P(O|\lambda_{AC}, AC_j)$ indicates the probability of the new class AC_j, given the model λ_{AC} and the observation sequence O. For the case of generative model Ω, this decision rule is expressed by the following equation

$$\begin{aligned} i &= \underset{j}{\boldsymbol{argmin}} P\left(O|\Omega_{AC_j}\right) \\ AC_j &= I_N - \{C_j\} \\ I_N &= \{C_1, C_2, \ldots, C_N\}, \end{aligned} \tag{3}$$

where i is the recognized class, Ω_{ACj} is the model which has been trained with the new class AC_j, which will hereafter be referred to as *complementary generative model*, and $P(O|\Omega_{ACj})$ indicates the output probability of model Ω_{ACj}, given the observation sequence O.

Figure 2 schematically depicts the training procedure for a discriminative model λ and its corresponding complementary model λ_{AC}. In Fig. 3, the training procedure for a generative model Ω and its corresponding complementary

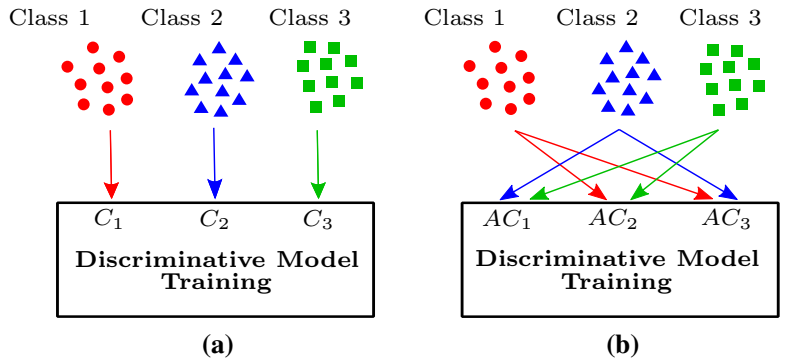


Fig. 2 Training procedure of the proposed complementary models for the case of discriminative models and a vocabulary of size N = 3. **a** Traditional model with classes $\{C_1, C_2, C_3\}$. **b** Complementary model with the new set of classes $\{AC_1, AC_2, AC_3\}$

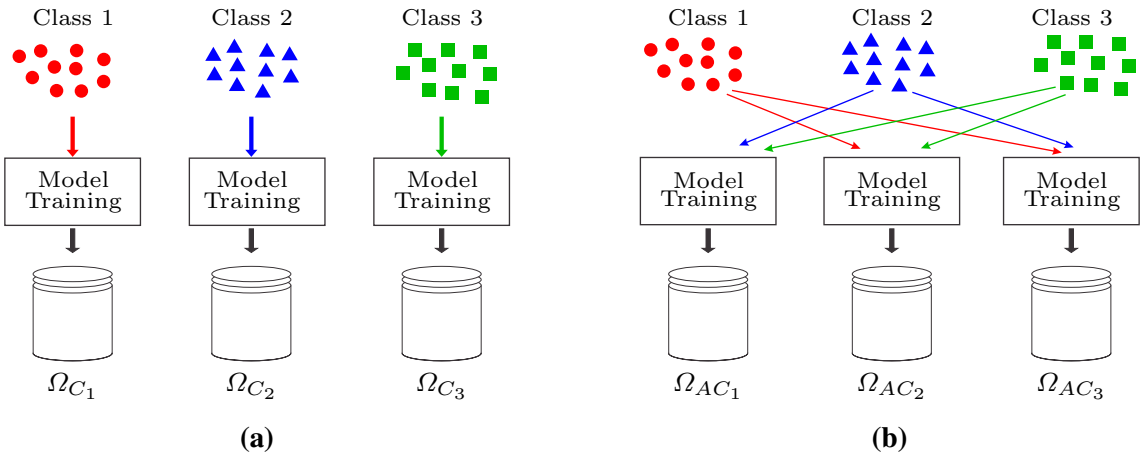


Fig. 3 Training procedure of the proposed complementary models for the case of generative models and a vocabulary of size N = 3. **a** Traditional models $\{\Omega_{C1}, \Omega_{C2}, \Omega_{C3}\}$, one for each class. **b** Complementary models $\{\Omega_{AC1}, \Omega_{AC2}, \Omega_{AC3}\}$, one for each new class $\{AC_1, AC_2, AC_3\}$

model Ω_{AC}, is also shown. In both cases, the vocabulary has three different classes (N = 3).

4 Proposed cascade of classifiers

In this section, a cascade classification structure is proposed to improve recognition rates. This scheme for speech classification is based on the combination of traditional and complementary models. The proposed combination scheme uses the complementary models to rescore only when it is believed that the traditional system is incorrect. Also, it can be employed with different kinds of input information, viz., audio, visual or audio-visual information, indistinctly. Given a model, the proposed cascade of classifiers is formed with the traditional model and the complementary version of the traditional model, based on the method proposed in Sect. 3. A schematic representation of the proposed speech classification scheme is depicted in Fig. 4.

In the training stage, the features obtained from the input audio-visual speech data are used to train the traditional model (λ or Ω). This model is used to detect the conflicting classes, based on the analysis of the confusion matrix described in Sect. 2. Based on this analysis, the complementary models are generated and trained. The test stage is composed by two steps. First, a classification based on the traditional model is performed, and the M most likely classes are pre-selected using the output probabilities. At this point, the possible solutions are reduced to these M classes. If none of these pre-selected classes are considered as conflicted ones (based on the analysis carried out in the training stage), the recognized word corresponds to the one with the highest probability. Otherwise, if any pair of these M classes is considered as conflicted ones, a second step is performed. In this step, the complementary model (λ_{AC} or Ω_{AC}) associated with these M classes is selected and used to determine the recognized class.

Figure 5 schematically depicts the cascade of classifiers scheme proposed in this paper, for the case of M = 3

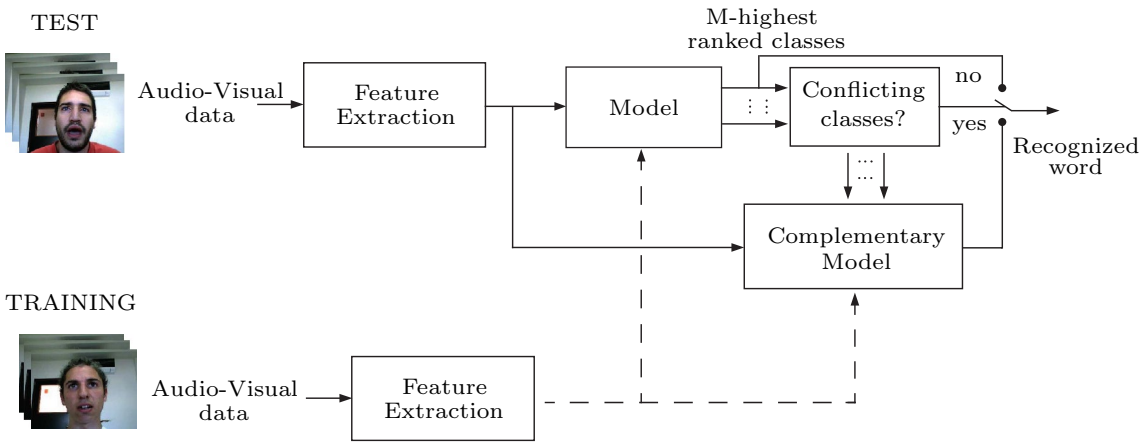


Fig. 4 Schematic representation of the proposed cascade of classifiers based on complementary models for audio visual speech classification. The M-highest ranked classes are obtained by sorting the model's output probabilities

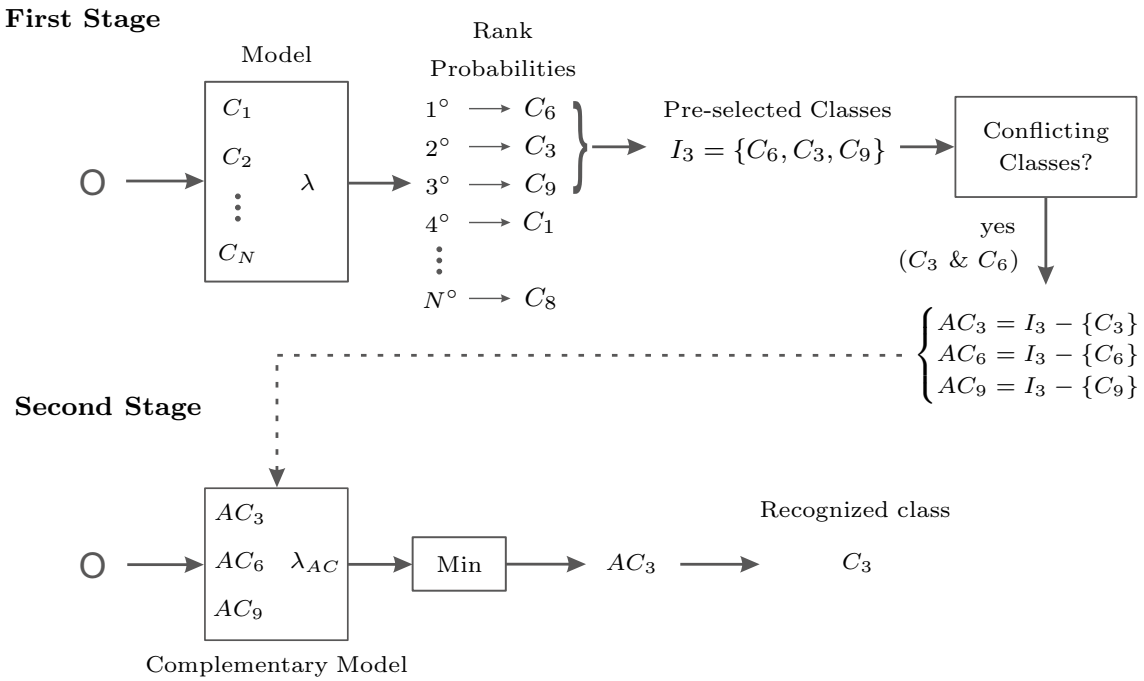


Fig. 5 Example of the proposed classifier combination strategy with M = 3 and implemented with a discriminative model λ. In the second stage, the complementary model (λ_{AC}) uses the new set of classes: {AC_3, AC_6, AC_9}

and implemented with a discriminative model λ. Given an observation sequence O, associated with the word to be recognized, the λ original model is evaluated, and its output probabilities (P(O|λ, C_i), i = 1, 2, …, N) are ranked. The M = 3 highest ranked values define the classes (I_3 = {C_6, C_3, C_9}) to be used for selecting the complementary model λ_{AC} (which has been previously trained in the training stage), because there are conflicting classes (C_3 and C_6). In this case, the new classes defined inside the complementary model are: $AC_3 = I_3 - \{C_3\}$ which is trained with the training data corresponding to classes C_6 and C_9, $AC_6 = I_3 - \{C_6\}$ which is trained with the training data corresponding to classes C_3 and C_9 and $AC_9 = I_3 - \{C_9\}$ which is trained with the training data corresponding to classes C_3 and C_6. Finally, the complementary model defined with these new classes is evaluated and the recognized word corresponds to the new class AC_i who gives the minimum probability.

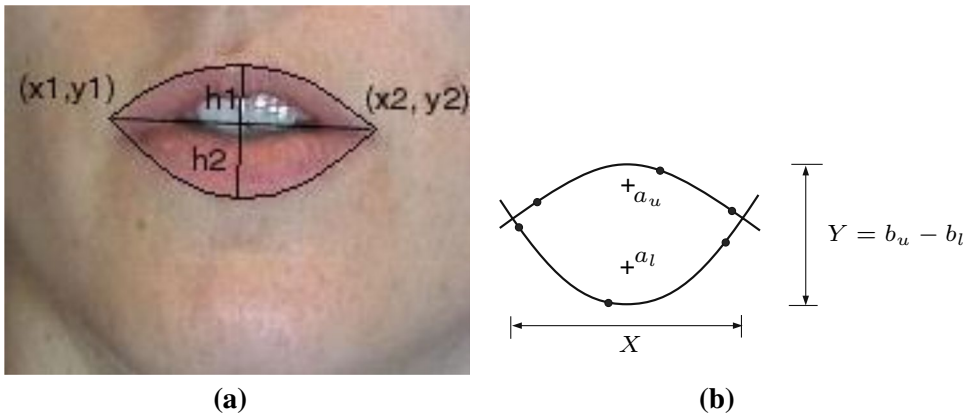


Fig. 6 AV-CMU database. **a** Visual data included in the database. **b** Parabolic lip contour model proposed in (Borgström & Alwan, 2008)

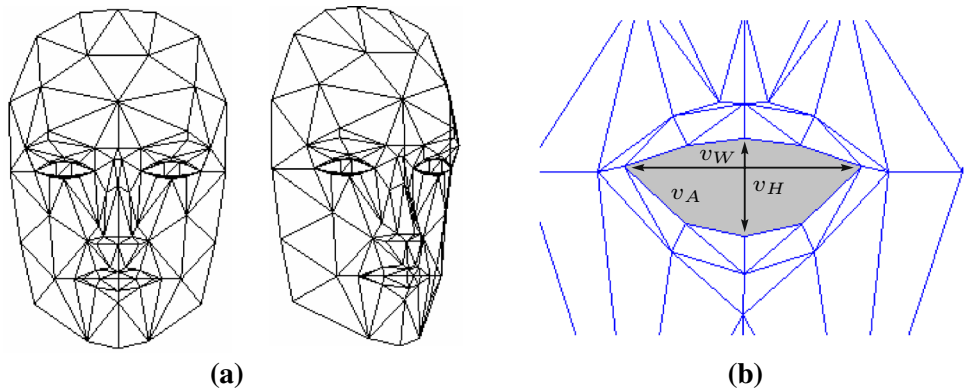


Fig. 7 AV-UNR Database visual features. **a** Candide-3 face model. **b** Visual parameters

5 Audio-visual databases

The performance of the proposed classification scheme is evaluated over three different audio-visual databases, two well-known public ones and another compiled by the authors of this paper. Each database has its own visual speech representation. In particular, for two databases visual features were computed using model-based methods, while image-based features were considered as visual information for the remaining one. Each database is described below.

5.1 AV-CMU database

The AV-CMU database (Huang & Chen, 1998) consists in the recording of ten speakers uttering a series of words. In this paper, a subset of ten words, corresponding to the numbers from 1 to 10 is considered for the experiments. These numbers were pronounced ten times by each speaker. Acoustic data was captured at 44.1 kHz, while visual data was captured at 30 frames per seconds. As depicted in Fig. 6a, this database provides the visual data through the position of mouth and lips at each video frame. In this paper, the parabolic lip contour model proposed in Borgström and Alwan (2008) is employed to represent the visual information during speech. This parabolic lip model, depicted in Fig. 6b, is fitted using the left (x_1, y_1) and the right (x_2, y_2) corners position of the mouth, and the heights of the openings of the upper (h_1) and lower (h_2) lips. Visual features are then represented by 5 parameters, viz., the focal parameters of the upper and lower parabolas, mouths width and height, and the main angle of the bounding rectangle of the mouth.

5.2 AV-UNR database

This database,[1] compiled by the authors of this paper, is composed by the utterances of a set of ten words (corresponding to actions such as open, close, save, stop, etc.), performed in random order by 16 speakers. Each word was pronounced 20 times by each speaker, resulting in a total of 3200 utterances. Audio data was captures at 8 kHz, while visual data was captured at 60 frames per seconds with a resolution of 640×480 pixels.

To represent the visual information during speech, the model-based method proposed in Terissi and Gómez (2010) is employed. This method extracts mouth visual features

[1] Available on the website: https://www.cifasis-conicet.gov.ar/AVdatabase/

Fig. 8 Example images from ten speakers of the AVLetters database

using a simple 3D face model, namely Candide-3 (Ahlberg, 2001). As depicted in Fig. 7, at each video frame, visual information is represented by 3 parameters, viz., mouth height (v_H), mouth width (v_W) and area between lips (v_A).

5.3 AVLetters database

The AVLetters database (Matthews et al., 2002) is composed by the recording of the letters (A–Z) by 10 speakers (5 males and 5 females). In these recordings, each speaker pronounced 3 times each letter, resulting in a total of 780 utterances. The original acoustic voice signals are not provided by this database. However, it includes the corresponding Mel-Frequency Cepstral Coefficients (MFCC) computed for each utterance. Visual data consists of mouth region images with a 80×60 pixels resolution, captured at 25 frames per seconds. Example images from the ten speakers are included in Fig. 8. To represent the visual information during speech, the method based on local spatiotemporal descriptors proposed in Zhao et al. (2009) is employed in this paper. This image-based method is applied directly over the image sequences of the AVLetters database. As a result of this method (Zhao et al., 2009), each image of the database is represented by a feature vector of 1770 coefficients.

6 Experimental results

In order to evaluate the performance and robustness of the proposed method in this paper, several experiments are carried out. As mentioned before, the cascade of classifiers proposed in Sect. 4 can handle different kinds of input features, so experiments with audio features (audio speech recognition), visual features (visual speech recognition or lip-reading) and fused audio-visual features (audio-visual speech recognition) are performed. Regarding the models and classifiers used to implement each stage of the system described in Sect. 4, four different cases are evaluated: Hidden Markov Model (HMM), Random Forest (RF), Support Vector Machine (SVM) and Adaptive Boosting (ADA) classifier. That is, a total of three discriminative models and one generative model are analyzed. The three databases reported in Sect. 5 are employed to carry out all the experiments described above.

There exist different factors in real applications that usually are not taken into account in the training stage of speech recognition systems. In most cases this is due to the complexity of the models or the lack of data representing all the possible scenarios in real situations. For example, visual occlusion in visual or audio-visual speech recognition, and acoustic noise affecting the audio information in audio or audio-visual speech recognition. Since the presence of acoustic noise is one of the most common sources of errors in speech recognition systems, experiments with noisy audio conditions are performed. Additive Gaussian and additive Babble noise (Krishnamurthy & Hansen, 2009) is injected in the acoustic channel for all the experiments where acoustic information is employed to extract the input features to the proposed system, in order to evaluate the robustness of the proposed cascade of classifiers. The NOISEX-92 database (Varga & Steeneken, 1993) is employed to extract samples of Babble noise, which is one of the most challenging noise conditions because is composed by speech from other speakers interfering the original speech to be recognized. This acoustic noise robustness analysis is carried out using clean audio information for the training stage and noise corrupted audio information (with signal-to-noise ratios (SNRs) ranging from $-$ 10 dB to 40 dB) for the testing stage of the proposed system.

In all the scenarios described above, a D-fold cross-validation (CV) procedure is carried out in order to obtain statistically significant recognition rates. Particularly, speaker independent evaluations are performed, using only one speaker in the testing stage and the remaining ones in the training stage, resulting in a tenfold for AV-CMU database and 16-fold for AV-UNR database. In order to fairly compare

the results obtained in this work over the AVLetters database with others methods reported in the literature, the evaluation protocol used by the authors in Hu et al. (2016); Matthews et al., 2002; Ngiam et al., 2011; Zhao et al., 2009) is employed. The first two utterances of each speaker are used in the training stage and the third utterance of each speaker is used in the testing stage. Unlike the case of AV-CMU and AV-UNR databases, for AVLetters database the evaluation is speaker dependent.

A frame wise analysis over the acoustic signal is performed and the first eleven non-DC Mel-Cepstral coefficients are extracted frame by frame. Also, their associated first and second time derivatives are calculated, which gives an audio feature vector composed by 33 parameters for each frame. In the audio-visual speech recognition experiments, the same frame wise analysis is performed, partitioning the acoustic signal in frames with the frame rate defined by the video channel frame rate. For each frame, the audio-visual feature vector is composed by the concatenation of the corresponding acoustic and visual features. This results in a fused audio-visual feature vector composed by 38 and 36 parameters for the AV-CMU and AV-UNR databases, respectively. Since the original recordings of the acoustic signals from AVLetters database are not available, experiments under noisy acoustic conditions for audio and audio-visual speech recognition scenarios could not be performed.

As described in Sect. 1, there are many models and classifiers proposed in the literature to perform classification and recognition tasks. The vast majority, especially when using discriminative models, cannot handle variable length input data, as is the case of speech recognition systems, due to its time varying nature. Each utterance, even of the same word, is very unlikely to have the same temporal duration. This leads to the need of some kind of normalization process to obtain fixed-length input data representation. To this end, the proposed method in Terissi et al. (2015a) based on a wavelet feature extraction technique, is employed in all the experiments with RF, SVM and ADA classifiers. This method requires three parameters to be defined, viz., the mother wavelet, the resolution level for the approximation and the normalized length of the resampled time functions. In this works, the wavelet resolution level was set to 3, the normalized length of the resampled time functions was set to 256, and the Daubechies 4 wavelet (db4) was chosen as the mother wavelet.

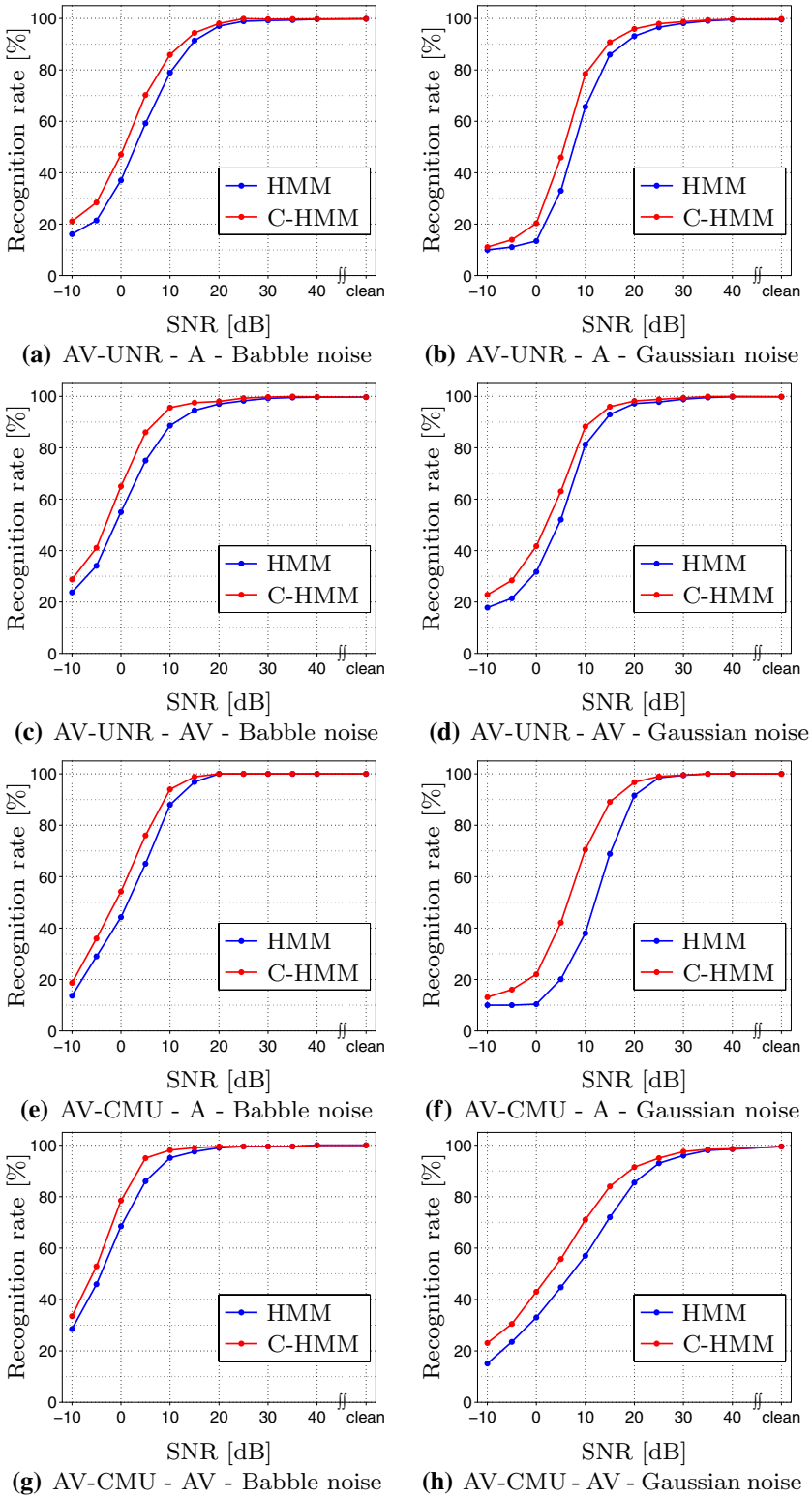
In order to disambiguate conflicting classification results, the conflicted classes must be recognized as a first step. The method proposed in Sect. 2 is employed, evaluating the classifier and selecting the 4 most conflicting pair of classes (P = 4) based on the analysis of the confusion matrix obtained. For each scenario, this evaluation is carried out randomly partitioning the whole data of each database in a training set (70%) and a test set (30%). These 4 pairs of conflicted classes are used in the second stage of the proposed system in Fig. 4 in order to determine if there are conflicting results for the observation being processed in the testing stage. The complementary models represented in Fig. 4 were implemented as 3-class complementary (M = 3).

6.1 Hidden Markov models

In the first stage of the proposed scheme, the classifiers are implemented with HMM (Rabiner, 1989) while in the second stage the complementary models are implemented with Gaussian Mixture Models (GMMs), namely Complementary Gaussian Mixture Models (CGMMs). The classical N-state left-to-right structure with continuous symbol observation and a linear combination of M_h Gaussian distributions as representation for each state, was employed for the HMMs implementation. The GMMs used to implement the CGMMs were modeled by a linear combination of M_g Gaussian distributions with continuous observations. The well-known Expectation–Maximization (EM) algorithm was employed for training both, the HMMs and the GMMs.

In this case, there are three tuning parameters of the proposed system, namely, the number of states (N) and Gaussian distributions (M_h) of the HMMs, and the number of Gaussian distributions (M_g) of the CGMMs, which are optimized via exhaustive search. Several experiments were carried out using values of N in the range from 1 to 20, M_h from 1 to 30 and M_g from 8 to 256, looking for the combination giving the best recognition results.

The recognition rates obtained at different SNRs over the AV-CMU and AV-UNR databases, using audio-only information and fused audio-visual information are depicted in Fig. 9, for the case of HMM and the proposed cascade of classifiers (C-HMM). It is clear that, for both databases, both types of inputs and both kinds of acoustic noise, the proposed scheme (C-HMM) performs better than the ones based on HMMs. This is more notorious at low and middle SNRs. In some cases there are very significant improvements, for example in Fig. 9f, for the case of AV-CMU database using audio-only information and Gaussian noise, reaching almost 33% of improvement for SNR = 10 dB. The results for the lip-reading scenario for AV-UNR, AV-CMU and AVLetters databases, are shown in Table 1. As it can be observed, the use of the proposed cascade of classifiers scheme (C-HMM) improves the recognition rates for all databases.



(a) AV-UNR - A - Babble noise **(b)** AV-UNR - A - Gaussian noise

(c) AV-UNR - AV - Babble noise **(d)** AV-UNR - AV - Gaussian noise

(e) AV-CMU - A - Babble noise **(f)** AV-CMU - A - Gaussian noise

(g) AV-CMU - AV - Babble noise **(h)** AV-CMU - AV - Gaussian noise

Fig. 9 Recognition based on HMM classifier and the proposed cascade of classifiers C-HMM. Recognition rates obtained over the AV-CMU and AV-UNR databases, using only acoustic (A) information and audio-visual (AV) information, for the cases of considering Babble noise (first column) and Gaussian noise (second column)

Table 1 Lip-reading based on HMM classifier

Database	Visual features	HMM (%)	C-HMM (%)
AV-UNR	Mouth shape parameters (Terissi & Gómez, 2010)	70.16	74.04
AV-CMU	Lip contour model (Borgström & Alwan, 2008)	57.79	61.82
AVLetters	Local spatiotemporal descriptors (Zhao et al., 2009)	57.30	61.25

Recognition rates based on HMM classifier and the proposed cascade of classifiers C-HMM over AV-UNR, AV-CMU and AVLeters

6.2 Random forest

In this case, both stages of the proposed cascade of classifiers are implemented with RF (Breiman, 2001). The complementary model based on RF will be hereafter referred to as Complementary Random Forest (CRF).

In this case, there are four parameters to adjust, namely, the number of trees in the ensemble (N) and the size of the random subset of features used at each splitting node in the tree (α), for both, the RFs and the CRFs. It is well known that if the number of trees in the ensemble is large enough (usually 500 or more), the particular value employed does not significantly affects the performance of the Random Forest classifier. In this paper, the value of N was set to 2000. In this way, finally there remain only two tuning parameters in the proposed system, namely, the size of the random subset of features used at each splitting node in the tree (α) for the RFs and the CRFs, which were optimized via exhaustive search using values in the range from 2 to 10.

Figure 10 shows the results obtained with RF and the proposed cascade of classifiers (C-RF), for the same experimental scenarios evaluated for the HMM case. Again, the proposed scheme (C-RF) yields better results than the RF-based scheme, which is more noticeable at low and medium SNRs. In this case, the maximum improvement achieved by resorting to the proposed cascade of classifiers is almost 16%, for the case of AV-UNR database using audio-only information and Gaussian noise with SNR = 0 dB. Table 2 shows the results obtained for the lip-reading case over all the databases employed. Again, the proposed scheme (C-RF) gives better results for all cases. In this case, a significant improvement of 7% for AVLetters database is achieved.

6.3 Support vector machine

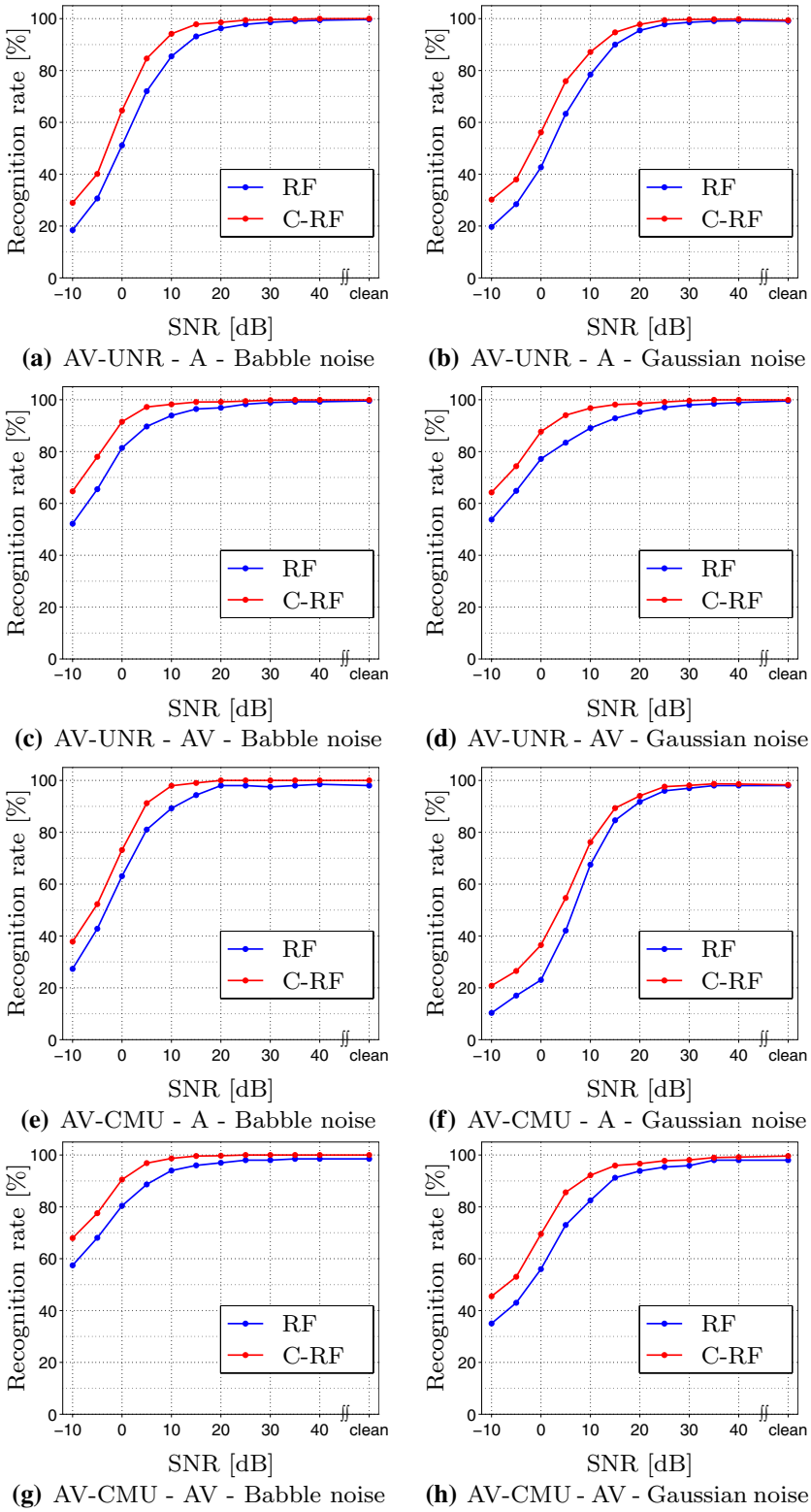
Simlar to the RF case, both stages of the cascade of classifiers proposed are implemented with SVM (Cortes & Vapnik, 1995), based on Gaussian kernels. The complementary model based on SVM will be hereafter referred to as Complementary Support Vector Machines (CSVM). The tuning parameters are four in this case, namely, the cost C and the σ value of the Gaussian kernel, for both classifiers SVM and CSVM, which were optimized via exhaustive search in a two stages procedure. First, a rough search was carried out, varying the C and σ values in decade steps ([…, 10^{-2}, 10^{-1}, 1, 10^{1}, 10^{2}, …]). Finally, a second finest search was carried out, using smaller steps for the C and σ values, where the best performance of the recognition system was achieved in the C/σ parameters space at the first stage.

The results obtained with SVM and the proposed cascade of classifiers (C-SVM) under the same scenarios as before are shown in Fig. 11. As can be seen, a performance improvement is achieved using the proposed scheme (C-SVM), which is more significant at low and medium SNRs. In this case, the proposed scheme performs better for the case of Babble noise in comparison with the case of Gaussian noise. The maximum improvement achieved by using the proposed cascade of classifiers is almost 13%, for the case of AV-CMU database using fused audio-visual information and Gaussian noise with SNR = 0 dB. In Table 3, the results for the lip-reading case over the three databases are depicted. As in the case of C-HMM and C-RF, significant improvements in the results for the three databases are obtained when using the proposed scheme (C-SVM), reaching a 7% improvement for the AVLetters database.

6.4 AdaBoost

Again, both stages of the cascade of classifiers proposed are implemented using Adaptive Boosting (ADA) (Schapire & Singer, 1999). The complementary model based on ADA will be hereafter referred to as Complementary AdaBoost (CADA). In this case, there are four tuning parameters, namely, the depth of each tree in the ensemble (d) and the number of iterations of the boosting algorithm (N), for both, the ADA and the CADA classifiers, which were optimized via exhaustive search. Several experiments were carried out using values of d in the range from 2 to 20, and considering N = [100, 500, 1000, 2000, 5000], looking for the best combination.

As in the previous subsections, the recognition rates obtained with ADA and the proposed scheme (C-ADA)



(a) AV-UNR - A - Babble noise **(b)** AV-UNR - A - Gaussian noise

(c) AV-UNR - AV - Babble noise **(d)** AV-UNR - AV - Gaussian noise

(e) AV-CMU - A - Babble noise **(f)** AV-CMU - A - Gaussian noise

(g) AV-CMU - AV - Babble noise **(h)** AV-CMU - AV - Gaussian noise

Fig. 10 Recognition based on RF classifier and the proposed cascade of classifiers C-RF. Recognition rates obtained over the AV-CMU and AV-UNR databases, using only acoustic (A) information and audio-visual (AV) information, for the cases of considering Babble noise (first column) and Gaussian noise (second column)

Table 2 Lip-reading based on RF classifier

Database	Visual features	RF (%)	C-RF
AV-UNR	Mouth shape parameters (Terissi & Gómez, 2010)]]	88.67	91.97
AV-CMU	Lip contour model (Borgström & Alwan, 2008)	71.65	75.05
AVLetters	Local spatiotemporal descriptors (Zhao et al., 2009)	65.38	72.34

Recognition rates based on RF classifier and the proposed cascade of classifiers C-RF over AV-UNR, AV-CMU and AVLeters

Table 3 Lip-reading based on SVM classifier

Database	Visual features	SVM (%)	C-SVM (%)
AV-UNR	Mouth shape parameters (Terissi & Gómez, 2010)	83.75	88.24
AV-CMU	Lip contour model (Borgström & Alwan, 2008)	67.53	73.15
AVLetters	Local spatiotemporal descriptors (Zhao et al., 2009)	63.08	69.97

Recognition rates based on SVM classifier and the proposed cascade of classifiers C-SVM over AV-UNR, AV-CMU and AVLeters

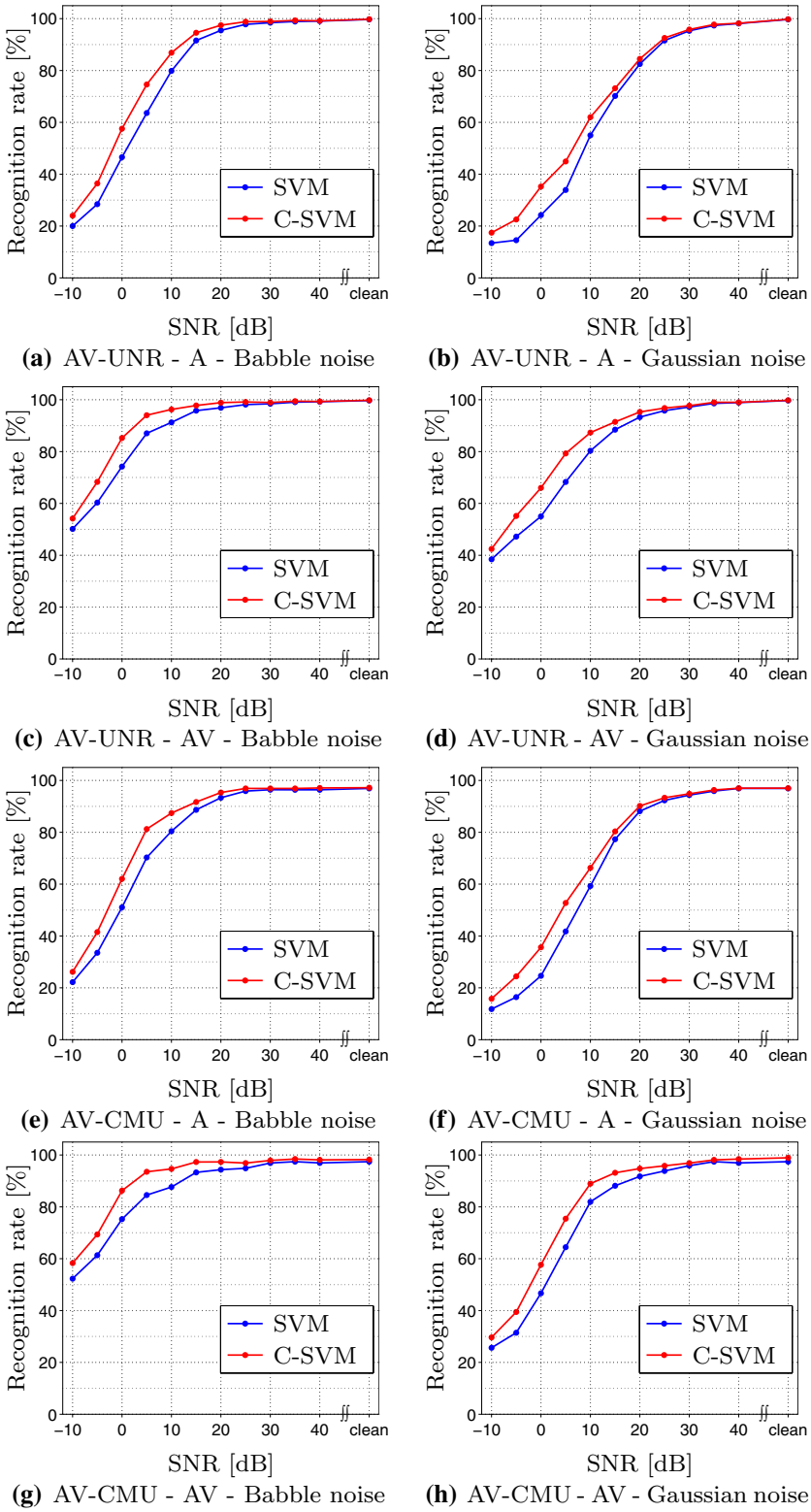
over the same experimental scenarios are shown in Fig. 12. As in the previous cases, an improvement in recognition rates is achieved when using the proposed cascade scheme (C-ADA) in all the evaluated scenarios. Again, this improvement is more noticeable at middle SNRs. The maximum improvement by resorting to the proposed cascade of classifiers is almost 13%, for the case of AV-CMU database using fused audio-visual information and Gaussian noise with SNR = 15 dB. Recognition rates obtained for the lip-reading scenario in all the databases used are depicted in Table 4. As in the three previous cases, better results are achieved by resorting to the proposed scheme (C-ADA) for all databases employed, with a maximum improvement of 6.3% for the AVLetters database.

6.5 Analysis and comparison

As it can be observed from the above presented results, the proposed cascade of classifiers based on complementary models achieved recognition rate improvements in comparison to the case of using only the traditional model, regardless of the experimental scenario setup (model/database/input information/kind of noise). This is more notorious at low and middle range of SNRs. Comparing the results obtained for each of the models used (C-HMM, C-RF, C-SVM and C-ADA), it can be observed that in general the improvements obtained were greater for the case of C-HMM and C-RF, reaching very significant improvements in some cases (33% for C-HMM). Apart from the magnitude of the improvements obtained in the recognition rates by resorting to the proposed scheme, it is interesting to note that in none of the experiments carried out, the results obtained by the cascade of classifiers were worse than that obtained with the traditional model. The cascade strategy proposed in this paper could be thought as a block that is added to the output of the traditional model. That is, the proposed strategy could be applied to obtain recognition rate improvements on an existing model simply by training the complementary models, without any modification on the traditional models.

In Table 5, the recognition rates obtained for the lip-reading scenario over the AVLetters database by the proposed C-HMM, C-RF, C-SVM and C-ADA schemes are presented. Also, the results obtained with other methods proposed in the literature are shown. As it can be observed, the proposed cascade of classifiers based on complementary models performs satisfactorily and better than the other approaches, independently of the model being used. In Fig. 13, the confusion matrix obtained in the experiments with the proposed C-RF scheme for the lip-reading scenario over the AVLetters database is shown. Also, the confusion matrix obtained with the RF classifier for the same experimental setup is shown. As it can be observed in Fig. 13a, there exist some conflicting classes in the RF classifier: A-N, Q-U, S-X, B-P, D-T and L-N.

For this experiment, the conflicting classes selected in the C-RF scheme are A-N, Q-U, S-X, B-P, which corresponds to



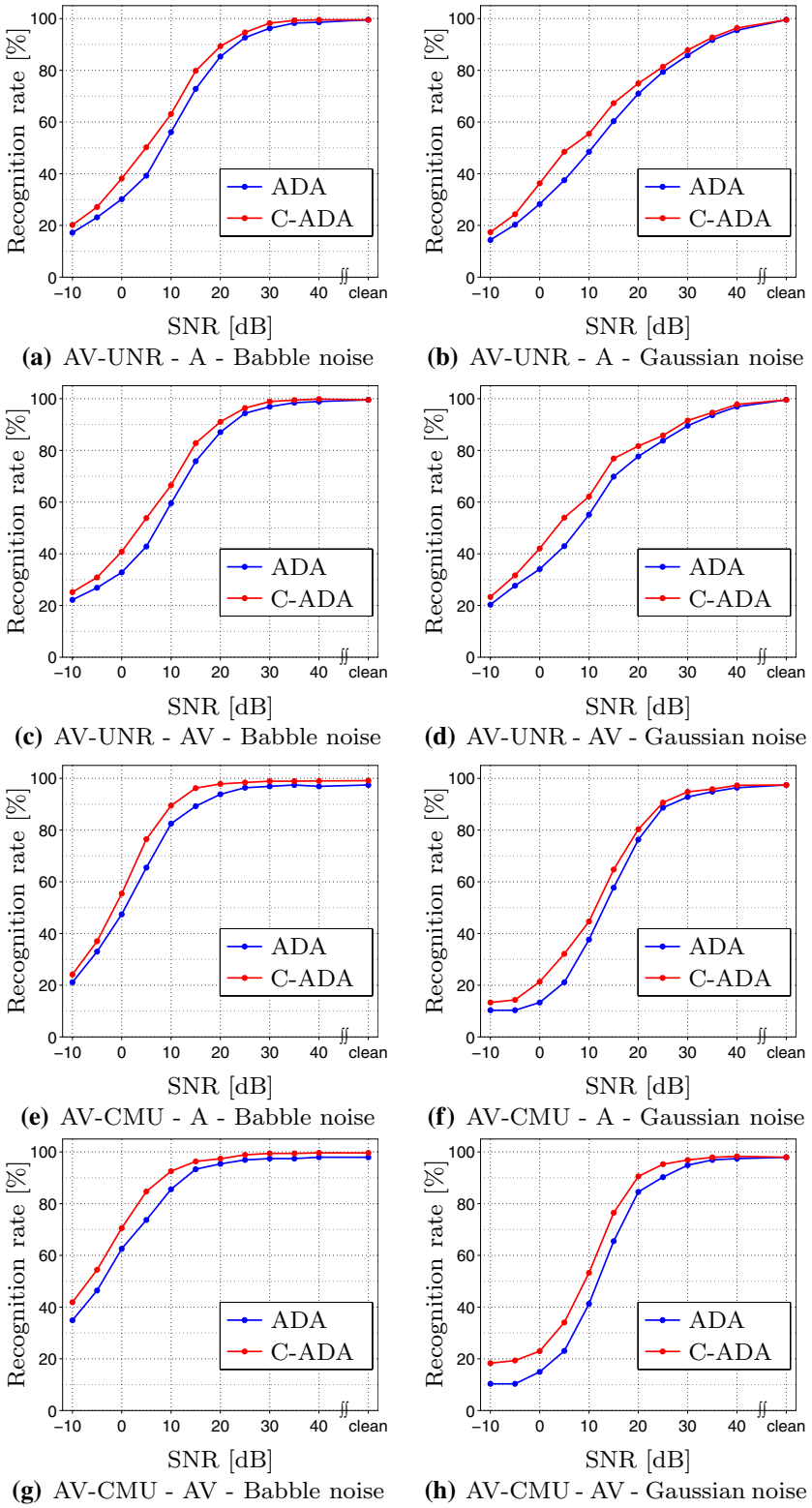
(a) AV-UNR - A - Babble noise **(b)** AV-UNR - A - Gaussian noise

(c) AV-UNR - AV - Babble noise **(d)** AV-UNR - AV - Gaussian noise

(e) AV-CMU - A - Babble noise **(f)** AV-CMU - A - Gaussian noise

(g) AV-CMU - AV - Babble noise **(h)** AV-CMU - AV - Gaussian noise

Fig. 11 Recognition based on SVM classifier and the proposed cascade of classifiers C-SVM. Recognition rates obtained over the AV-CMU and AV-UNR databases, using only acoustic (A) information and audio-visual (AV) information, for the cases of considering Babble noise (first column) and Gaussian noise (second column)



(a) AV-UNR - A - Babble noise **(b)** AV-UNR - A - Gaussian noise

(c) AV-UNR - AV - Babble noise **(d)** AV-UNR - AV - Gaussian noise

(e) AV-CMU - A - Babble noise **(f)** AV-CMU - A - Gaussian noise

(g) AV-CMU - AV - Babble noise **(h)** AV-CMU - AV - Gaussian noise

Fig. 12 Recognition based on ADA classifier and the proposed cascade of classifiers C-ADA. Recognition rates obtained over the AV-CMU and AV-UNR databases, using only acoustic (A) information and audio-visual (AV) information, for the cases of considering Babble noise (first column) and Gaussian noise (second column)

Table 4 Lip-reading based on ADA classifier

Database	Visual features	ADA (%)	C-ADA (%)
AV-UNR	Mouth shape parameters (Terissi & Gómez, 2010)	85.63	90.65
AV-CMU	Lip contour model (Borgström & Alwan, 2008)	67.53	72.24
AVLetters	Local spatiotemporal descriptors (Zhao et al., 2009)	54.23	60.57

Recognition rates based on ADA classifier and the proposed cascade of classifiers C-ADA over AV-UNR, AV-CMU and AVLeters

Table 5 Lip-reading over AVLetters database

Classifier	Visual features	Accuracy (%)
C-HMM	Local spatiotemporal descriptors (Zhao et al., 2009)	61.25
C-RF	**Local spatiotemporal descriptors (Zhao et al., 2009)**	**72.34**
C-SVM	Local spatiotemporal descriptors (Zhao et al., 2009)	69.57
C-ADA	Local spatiotemporal descriptors (Zhao et al., 2009)	60.57
HMM (Matthews et al., 2002)	Multiscale spatial analysis (Matthews et al., 2002)	44.60
HMM (Zhao et al., 2009)	Local spatiotemporal descriptors (Zhao et al., 2009)	57.30
SVM (Zhao et al., 2009)	Local spatiotemporal descriptors (Zhao et al., 2009)	58.85
Deep Autoencoder (Ngiam et al., 2011)	Video-only deep autoencoder (Ngiam et al., 2011)	64.40
RTMRBM (Hu et al., 2016)	Mouth region PCA (Hu et al., 2016)	64.63

Recognition rates obtained with the proposed cascade of classifiers for the case of HMM (C-HMM), RF (C-RF), SVM (C-SVM), ADA (C-ADA), and also by others methods in the literature

some of the conflicting classes actually observed in the RF classifier's experiments. As it can be seen in the confusion matrix depicted in Fig. 13b, the objective of disambiguating conflicting classes is achieved.

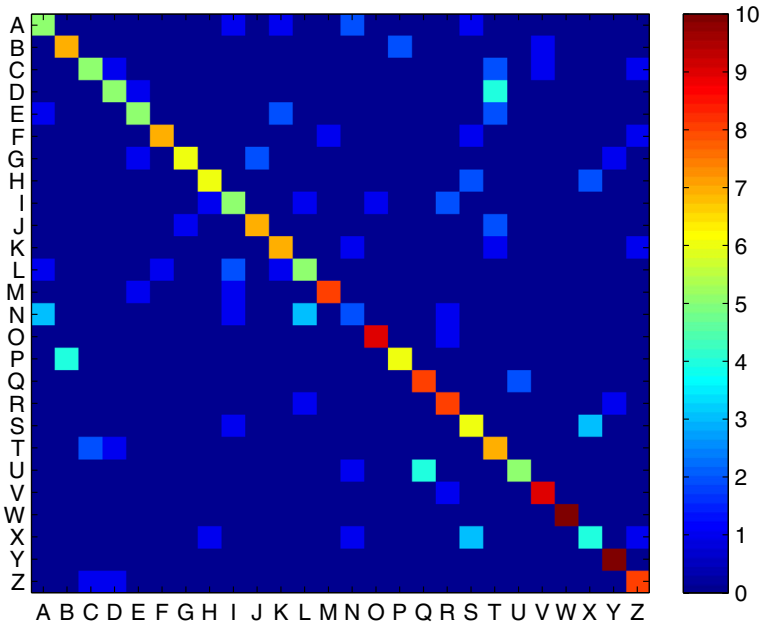
In Table 6, the recognition rates obtained for the lip-reading scenario over the AV-CMU database by the proposed C-HMM, C-RF, C-SVM and C-ADA systems, are depicted. The results obtained with other methods proposed in the literature are also shown. Again, the proposed cascade of classifiers based on complementary models performs satisfactorily and better than the other approaches, independently of the model being used.

7 Future works

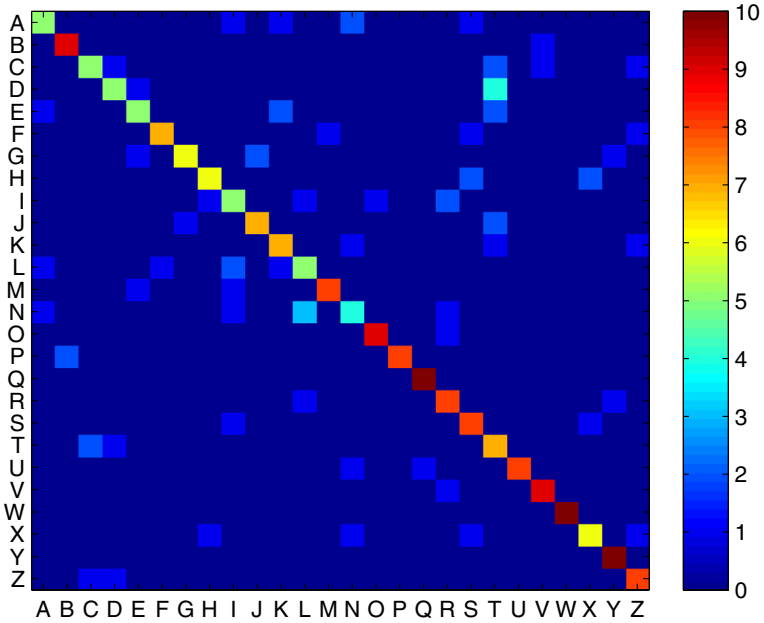
As can be seen in Sect. 6, the value of the parameter P, which represents the number of pairs of conflicting classes that will define the complementary models proposed in this paper, was manually selected. This value was obtained after several experiments, selecting the value of P that maximized the recognition rate. Therefore, automatically selecting this value is the next natural step to improve the proposed scheme. Work is in progress with promising results.

8 Conclusions

This papers proposes a novel scheme for Audio Visual Speech Classification tasks based on the combination of traditional and complementary models in a cascade structure, to improve recognition rates. This proposed scheme can be employed either for different generative models or discriminative models. Also, it can handle different kinds of input information, viz., audio, visual or audio-visual information, indistinctly. The concept of Complementary Models is introduced, which is based in a novel training procedure. The main idea is to detect the absence of a class by redefining classes, and to this end, for each particular word in the vocabulary, a new class is defined using all the instances of the words in the vocabulary except the corresponding to the one being represented. Four different models, viz., Hidden Markov Models, Random Forest, Support Vector Machines and Adaptive Boosting, were employed to analyze the efficiency of the proposed cascade of classifiers. The performance of the proposed speech classification scheme was evaluated at different conditions, considering only audio information, only video information (lip-reading), and fused audio-visual information. These evaluations were carried out over three different audio-visual databases, two of them public ones and the remaining one compiled



(a) AVLetters - Lip-reading - RF classifier



(b) AVLetters - Lip-reading - proposed cascade of classifiers C-RF

Fig. 13 Lip-reading over AVLetters database. Confusion matrix of the recognition results based on **a** RF classifier and **b** the proposed cascade of classifiers C-RF

by the authors of this paper. The robustness of the proposed scheme against noisy conditions in the acoustic channel is also evaluated. Experimental results show that a good performance is achieved with the proposed system over the three databases and for the different kinds of input information being considered. Recognition rates improvements are achieved by means of the proposed cascade of classifiers in all the scenarios and for the four models used. In addition, the proposed method performs better than other reported methods in the literature over the same two public databases. The proposed strategy could be applied to obtain recognition rates improvements on an existing model simply by training the complementary models, without any modification on the traditional models.

Table 6 Lip-reading over AV-CMU database

Classifier	Visual features	Accuracy (%)
C-HMM	Lip contour model (Borgström & Alwan, 2008)	61.82
C-RF	**Lip contour model (Borgström & Alwan,** 2008)	**75.05**
C-SVM	Lip contour model (Borgström & Alwan, 2008)	73.15
C-ADA	Lip contour model (Borgström & Alwan, 2008)	72.24
HMM (Borgström & Alwan, 2008)	Lip contour model (Borgström & Alwan, 2008)	61.17

Recognition rates obtained with the proposed cascade of classifiers for the case of HMM (C-HMM), RF (C-RF), SVM (C-SVM), ADA (C-ADA), and also by other method in the literature

Author contribution Not applicable.

Funding This study was funded by Agencia Nacional de Promoción Científica y Tecnológica (Grant No.: PICT 2018-01802), Universidad Nacional de Rosario (Grant No.: PID-UNR - 80020190100233UR).

Data availability The AV-UNR database compiled by the authors of this paper its available on the website: https://www.cifasis-conicet.gov.ar/AVdatabase/.

Code availability Not applicable.

Declarations

Conflicts of interest Not applicable.

References

Aggarwal, R. K., & Dave, M. (2012). Integration of multiple acoustic and language models for improved Hindi speech recognition system. *International Journal of Speech Technology, 15*(2), 165–180.

Ahlberg, J. (2001). Candide-3—an updated parameterized face. *Technical report*, Department of Electrical Engineering, Linkping University, Sweden

Ahmadi, S., Ahadi, S. M., Cranen, B., & Boves, L. (2014). Sparse coding of the modulation spectrum for noise-robust automatic speech recognition. *EURASIP Journal on Audio, Speech, and Music Processing, 2014*(1), 36.

Amer, M. R., Siddiquie, B., Khan, S., Divakaran, A., Sawhney, H. (2014). Multimodal fusion using dynamic hybrid models. In: *Proceedings of IEEE winter conference on applications of computer vision*, pp 556–563

Benhaim, E., Sahbi, H., & Vitte, G. (2014). Continuous visual speech recognition for multimodal fusion. In: *Proceedings of IEEE international conference on acoustics, speech and signal processing*, pp. 4618–4622

Bilmes, J. A., & Kirchhoff, K. (2003). Generalized rules for combination and joint training of classifiers. *Pattern Analysis and Applications, 6*(3), 201–211.

Biswas, A., Sahu, P. K., & Chandra, M. (2016). Multiple cameras audio visual speech recognition using active appearance model visual features in car environment. *International Journal of Speech Technology, 19*(1), 159–171.

Borgström, B., & Alwan, A. (2008). 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, 38*(6), 1273–1280.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

Breslin, C. (2008). Generation and combination of complementary systems for automatic speech recognition. Ph.D. thesis, Cambridge University

Breslin, C., & Gales, M. (2009). Directed decision trees for generating complementary systems. *Speech Communication, 51*(3), 284–295.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273–297.

Deng, L., & Li, X. (2013). Machine learning paradigms for speech recognition: An overview. *IEEE Transactions on Audio, Speech, and Language Processing, 21*(5), 1060–1089.

Dietterich, T.G. (2000). Ensemble methods in machine learning. In: *Multiple classifier systems. Lecture notes in computer science* (vol. 1857, pp. 1–15). Berlin: Springer

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

Gales, M. J. F., Kim, D. Y., Woodland, P. C., Chan, H. Y., Mrva, D., Sinha, R., & Tranter, S. E. (2006). Progress in the CU-HTK broadcast news transcription system. *IEEE Transactions on Audio, Speech, and Language Processing, 14*(5), 1513–1525.

Guglani, J., & Mishra, A. N. (2020). DNN based continuous speech recognition system of Punjabi language on Kaldi toolkit. *International Journal of Speech Technology*. https://doi.org/10.1007/s10772-020-09717-8

Hain, T., Burget, L., Dines, J., Garau, G., Wan, V., Karafi, M., Vepa, J., & Lincoln, M. (2007). The AMI system for the transcription of speech in meetings. *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 4*, 357–360.

Hu, D., Li, X., & Lu, X. (2016). Temporal multimodal learning in audiovisual speech recognition. In: *Proceedings of IEEE conference on computer vision and pattern recognition*, pp. 3574–3582

Hu, R., & Zhao, Y. (2007). A bayesian approach for phonetic decision tree state tying in conversational speech recognition. *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 4*, 661–664.

Huang, F. J., & Chen, T. (1998). *Advanced multimedia processing laboratory*. Cornell University. Retrieved June 2020, from http://chenlab.ece.cornell.edu/projects/AudioVisualSpeechProcessing.

Hwang, M., Wang, W., Lei, X., Zheng, J., Cetin, O., & Peng, G. (2007). Advances in mandarin broadcast speech recognition. In: *Proceedings of the 8th annual conference of the international speech communication association*, pp. 2613–2616

Jaimes, A., & Sebe, N. (2007). Multimodal human-computer interaction: A survey. *Computer Vision and Image Understanding, 108*(1–2), 116–134.

Katsaggelos, A. K., Bahaadini, S., & Molina, R. (2015). Audiovisual fusion: Challenges and new approaches. *Proceedings of the IEEE, 103*(9), 1635–1653.

Kittler, J., Hatef, M., Duin, R. P. W., & Matas, J. (1998). On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*(3), 226–239.

Koziarski, M., Krawczyk, B., & Wozniak, M. (2017). The deterministic subspace method for constructing classifier ensembles. *Pattern Analysis and Applications, 20*(4), 981–990.

Krawczyk, B., & Cyganek, B. (2017). Selecting locally specialised classifiers for one-class classification ensembles. *Pattern Analysis and Applications, 20*(2), 427–439.

Krishnamurthy, N., & Hansen, J. (2009). Babble noise: Modeling, analysis, and applications. *IEEE Transactions on Audio, Speech, and Language Processing, 17*(7), 1394–1407.

Liu, C. L., Hao, H., & Sako, H. (2004). Confidence transformation for combining classifiers. *Pattern Analysis and Applications, 7*(1), 2–17.

Matthews, I., Cootes, T., Bangham, J. A., Cox, S., & Harvey, R. (2002). Extraction of visual features for lipreading. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 24*(2), 198–213.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*(5588), 746–748.

Najkar, N., Razzazi, F., & Sameti, H. (2014). An evolutionary decoding method for HMM-based continuous speech recognition systems using particle swarm optimization. *Pattern Analysis and Applications, 17*(2), 327–339.

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A. (2011). Multimodal deep learning. In: *Proceedings of the 28th international conference on machine learning*, pp. 689–696

Noda, K., Yamaguchi, Y., Nakadai, K., Okuno, H. G., & Ogata, T. (2015). Audio-visual speech recognition using deep learning. *Applied Intelligence, 42*(4), 722–737.

Pao, T., Liao, W., Wu, T., & Lin, C. (2009). Automatic visual feature extraction for Mandarin audio-visual speech recognition. In: *Proceedings of IEEE international conference on systems, man and cybernetics*, pp. 2936–2940

Papandreou, G., Katsamanis, A., Pitsikalis, V., & Maragos, P. (2009). Adaptive multimodal fusion by uncertainty compensation with application to audiovisual speech recognition. *Transactions on Audio, Speech, and Language Processing, 17*(3), 423–435.

Potamianos, G., Neti, C., Gravier, G., Garg, A., & Senior, A. W. (2003). Recent advances in the automatic recognition of audio-visual speech. *Proceedings of the IEEE, 91*(9), 1306–1326.

Potamianos, G., Neti, C., Iyengar, G., Senior, A. W., & Verma, A. (2001). A cascade visual front end for speaker independent automatic speechreading. *International Journal of Speech Technology, 4*(3), 193–208.

Prieto, O. J., Alonso-González, C. J., & Rodríguez, J. J. (2015). Stacking for multivariate time series classification. *Pattern Analysis and Applications, 18*(2), 297–312.

Puurula, A., & Van Compernolle, D. (2010). Dual stream speech recognition using articulatory syllable models. *International Journal of Speech Technology, 13*(4), 219–230.

Rabiner, L. (1989). A tutorial on Hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE, 77*(2), 257–286.

Sad G. D., Terissi L. D., Gómez J. C. (2017). Decision level fusion for audio-visual speech recognition in noisy conditions. In C. Beltrán-Castañón, I. Nyström, F. Famili (Eds.), *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. CIARP 2016. Lecture Notes in Computer Science* (vol. 10125). Cham: Springer. https://doi.org/10.1007/978-3-319-52277-7_44.

Savchenko, A. V., & YaI, K. (2014). About neural-network algorithms application in viseme classification problem with face video in audiovisual speech recognition systems. *Optical Memory and Neural Networks (information Optics), 23*(1), 34–42.

Schapire, R. E., & Singer, Y. (1999). Improved boosting algorithms using confidence-rated predictions. *Machine Learning, 37*(3), 297–336.

Shen, P., Tamura, S., & Hayamizu, S. (2014). Multistream sparse representation features for noise robust audio-visual speech recognition. *Acoustical Science and Technology, 35*(1), 17–27.

Shin, J., Lee, J., & Kim, D. (2011). Real-time lip reading system for isolated Korean word recognition. *Pattern Recognition, 44*(3), 559–571.

Shivappa, S., Trivedi, M., & Rao, B. (2010). Audiovisual information fusion in human computer interfaces and intelligent environments: A survey. *Proceedings of the IEEE, 98*(10), 1692–1715.

Stewart, D., Seymour, R., Pass, A., & Ming, J. (2014). Robust audio-visual speech recognition under noisy audio-video conditions. *IEEE Transactions on Cybernetics, 44*(2), 175–184.

Stüker, S., Fügen, C., Burger, S., & Wölfel, M. (2006). Cross-system adaptation and combination for continuous speech recognition: the influence of phoneme set and acoustic front-end. In: *Proceedings of the 9th international conference on spoken language processing (INTERSPEECH 2006 - ICSLP)*, pp. 521–524

Tao, F., & Busso, C. (2020). End-to-end audiovisual speech recognition system with multitask learning. *IEEE Transactions on Multimedia*. https://doi.org/10.1109/TMM.2020.2975922

Terissi, L. D., & Gómez, J. C. (2010). 3D head pose and facial expression tracking using a single camera. *Journal of Universal Computer Science, 16*(6), 903–920.

Terissi, L. D., Sad, G. D., & Gómez, J. C. (2018). Robust front-end for audio, visual and audio-visual speech classification. *International Journal of Speech Technology, 21*(2), 293–307.

Terissi, L. D., Sad, G. D., Gómez, J. C., & Parodi, M. (2015a). Audio-visual speech recognition scheme based on wavelets and random forests classification. In: Pardo, A., and Kittler, J. (Eds.), *Progress in pattern recognition, image analysis, computer vision, and applications. CIARP 2015. Lecture notes in computer science* (vol. 9423, pp. 567–574). Cham: Springer

Terissi L. D., Sad G. D., Gómez J. C., & Parodi M. (2015b). Noisy speech recognition based on combined audio-visual classifiers. In F. Schwenker, S. Scherer, L. P. Morency (Eds.), *Multimodal pattern recognition of social signals in human-computer-interaction. MPRSS 2014. Lecture Notes in Computer Science* (vol. 8869). Cham: Springer. https://doi.org/10.1007/978-3-319-14899-1_5.

Vallet, F., Essid, S., & Carrive, J. (2013). A multimodal approach to speaker diarization on TV talk-shows. *IEEE Transactions on Multimedia, 15*(3), 509–520.

Varga, A., & Steeneken, H. J. M. (1993). Assessment for automatic speech recognition II: NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems. *Speech Communication, 12*(3), 247–251.

Wagner, P., Malisz, Z., & Kopp, S. (2014). Gesture and speech in interaction: An overview. *Speech Communication, 57*, 209–232.

Zeiler, S., Nicheli, R., Ma, N., Brown, G. J., & Kolossa, D. (2016). Robust audiovisual speech recognition using noise-adaptive linear discriminant analysis. In: *Proceedings of IEEE international conference on acoustics, speech and signal processing*, pp. 2797–2801

Zhao, G., Barnard, M., & Pietikäinen, M. (2009). Lipreading with local spatiotemporal descriptors. *IEEE Transactions on Multimedia, 11*(7), 1254–1265.