Robust front-end for audio, visual and audio–visual speech classification

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

International Journal of Speech Technology

ISSN 1381-2416

Int J Speech Technol
DOI 10.1007/s10772-018-9504-y

Springer

Your article is protected by copyright and all rights are held exclusively by Springer Science+Business Media, LLC, part of Springer Nature. This e-offprint is for personal use only and shall not be self-archived in electronic repositories. If you wish to self-archive your article, please use the accepted manuscript version for posting on your own website. You may further deposit the accepted manuscript version in any repository, provided it is only made publicly available 12 months after official publication or later and provided acknowledgement is given to the original source of publication and a link is inserted to the published article on Springer's website. The link must be accompanied by the following text: "The final publication is available at link.springer.com”.

Springer

Author's personal copy

International Journal of Speech Technology
https://doi.org/10.1007/s10772-018-9504-y

Robust front-end for audio, visual and audio–visual speech classification

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

Received: 27 October 2017 / Accepted: 29 March 2018
© Springer Science+Business Media, LLC, part of Springer Nature 2018

Abstract
This paper proposes a robust front-end for speech classification which can be employed with acoustic, visual or audio–visual information, indistinctly. Wavelet multiresolution analysis is employed to represent temporal input data associated with speech information. These wavelet-based features are then used as inputs to a Random Forest classifier to perform the speech classification. The performance of the proposed speech classification scheme is evaluated in different scenarios, namely, considering only acoustic information, only visual information (lip-reading), and fused audio–visual information. These evaluations are carried out over three different audio–visual databases, two of them public ones and the remaining one compiled by the authors of this paper. 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. In addition, the proposed method performs better than other reported methods in the literature over the same two public databases. All the experiments were implemented using the same configuration parameters. These results also indicate that the proposed method performs satisfactorily, neither requiring the tuning of the wavelet decomposition parameters nor of the Random Forests classifier parameters, for each particular database and input modalities.

Keywords Audio–visual speech recognition · Wavelet decomposition · Random forests

1 Introduction

The development of Multimodal Human Computer Interfaces (HCIs), which imitate the way humans communicate with each other, has attracted the attention of numerous research groups worldwide in the last decades. Audio Visual Speech Recognition is a fundamental task in HCIs, where the acoustic and visual information (mouth movements, facial gestures, etc.) during speech are taken into account, mimicking communication among humans, which is multimodal in nature. The correlation between the acoustic and visual information during speech is essential, for instance, for the hearing impaired people, and is also important for normal listeners to improve the intelligibility of the speech signal in noisy environments. Several strategies have been proposed in the literature for audio–visual speech recognition (Shivappa et al. 2010; Papandreou et al. 2009; Potamianos et al. 2003), where improvements of the recognition rates are achieved by fusing audio and visual features related to speech. As expected, these improvements are more prominent when the audio channel is corrupted by noise, which is a usual situation in speech recognition applications.

To perform this task, different methods have been proposed in the literature in order to extract, represent, combine and model audio and visual information. Most of these approaches represent the acoustic information related to speech based on classical Mel-cepstrum analysis (Rabiner 1989), or some modification of them (Maganti and Matassoni 2014; Trottier et al. 2015; Panda and Nayak 2016; Uluskan et al. 2017). However, there exist a large variety of methods to represent visual information during speech. These methods can be roughly categorized into model-based (Borgström and Alwan 2008) and image-based (Zhao et al. 2009). In model-based approaches, the

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

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

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

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

Published online: 13 April 2018

Author's personal copy

visual information is represented in terms of geometrical data, such as contours of lips (Saitoh et al. 2008; Iwano et al. 2007; Wang et al. 2004), Active Appearance Models (AAM) (Biswas et al. 2016), shape of jaw and cheek (Aleksic et al. 2002), facial animation parameters (Yau et al. 2007) and mouth width, mouth opening, oral cavity area and oral cavity perimeter (Matthews et al. 2002). These methods commonly require accurate and reliable facial and lip feature detection and tracking. On the other hand, image-based approaches extract visual information directly from the pixel level data, mainly resorting to Principal Component Analysis (PCA) (Gowdy et al. 2004), Discrete Cosine Transform (DCT) (Potamianos et al. 1998), Linear Discriminant Analysis (LDA) (Potamianos et al. 2001), *Zernike* moments (Borde et al. 2015) and spatiotemporal coding schemes (Zhao et al. 2009), among other techniques. Regarding the combination of audio and visual information, the proposed methods can be classified according to the way that audio and visual information is combined (or fused), viz., feature level fusion, classifier level fusion and decision level fusion (Dupont and Luettin 2000). Finally, for recognizing a given sequence of audio or audio–visual features, several kinds of pattern recognition methods have been adopted in the literature. Probably, the most widely used are those based on traditional HMMs Dupont and Luettin (2000); Foo et al. (2004); Miki et al. (2014); Dong et al. (2005); Papandreou et al. (2009); Puviarasan and Palanivel (2011); Estellers et al. (2012), which statistically model transitions between the speech classes, and assume a class-dependent generative model for the observed features. In addition, several approaches based on Artificial Neural Networks (ANN) (Potamianos et al. 2003), Linear Discriminant Analysis, Support Vector Machine classifiers (SVM) (Zhao et al. 2009), matching methods utilizing dynamic programming, K-Nearest Neighbors (K-NN) algorithms (Shin et al. 2011), Deep Learning (Ngiam et al. 2011; Yin et al. 2015; Katsaggelos et al. 2015; Petridis and Pantic 2016), Restricted Boltzmann Machines (Amer et al. 2014; Hu et al. 2016), sparse coding (Wright et al. 2010; Ahmadi et al. 2014; Monaci et al. 2009; Shen et al. 2014), just to mention some, have been also proposed.

In general, a calibration stage to tune the parameters of the classifier is required by these audio–visual recognition systems, in order to obtain adequate performances in the recognition task. This calibration is often performed by testing different combinations of the classifier's tuning parameters, which is usually a time consuming procedure. In addition, the optimal values for the parameters could depend on the particular audio–visual dataset being employed.

In this paper, a novel front-end for speech classification, which can be employed with audio, visual or audio–visual information, indistinctly, is proposed. This approach is based on wavelets and Random Forests (RF) Breiman (2001). The sequences of audio and visual parameters are represented in a compact form in terms of Wavelet multiresolution analysis. These wavelet-based features are then used as inputs to a Random Forest classifier to perform the speech recognition. The good characteristics of RF, such as very good discriminative capabilities, computational efficiency over large databases and the capability of handling thousands of input variables avoiding the need for variable selection, are inherited by the proposed speech recognition scheme.

Discrete Wavelet Transform (DWT) has already been used by others authors for speech recognition tasks. For instance, DWT was employed for denoising, applied as a preprocessing stage before feature extraction to compensate noise effects (Gowdy and Tufekci 2000; Farooq and Datta 2003b). Also, several parameterizations methods based on the DWT for robust automatic speech recognition have been proposed in the literature (Farooq and Datta 2003a; Gupta and Gilbert 2001; Kotnik et al. 2003; Pavez and Silva 2012; Tufekci et al. 2006), where the DWT coefficients are used to represent the speech signal, instead of the traditional Mel-Frequency Cepstrum Coefficients (MFCC). For example, in Ali et al. (2014) DWT is applied on the acoustic speech signal and its associated coefficients are used for word classification based on Linear Discriminant Analysis. In Rajeswari et al. (2014) a frontend is introduced which uses wavelets for both enhancement and feature extraction stages. On the other hand, in the current work DWT is proposed with the purpose of representing in a compact form the input sequences of acoustic and/or visual parameters associated with speech. Here, DWT is neither used for denoising nor for feature extraction, it is employed as a generic feature representation tool. On the other hand, very few works in the literature make use of Random Forests for speech classification, for instance, in Ali et al. (2015) and Attar et al. (2010), RF have been proposed for isolated word classification tasks. In these works, acoustic signal is represented in terms of its associated MFCC coefficients. To the best of the present authors' knowledge the proposed combination of DWT and RF for audio, visual, and fused audio–visual speech classification has not bee used in the literature before.

The proposed speech classification scheme can handle different kinds of input data. For that reason, its performance is evaluated in different scenarios, viz., considering only audio information, only video information (lip-reading), and fused audio–visual information, respectively. These evaluations are carried out over three different databases, viz., AVLetters database (Matthews et al. 2002), Carnegie Mellon University (CMU) database (Huang and Chen 1998), and a database compiled by the authors, hereafter referred to as AV-1 database (no further information about this database is provided at these stage to compliant the double-blind reviewing process). In particular, in AV-CMU and AV-1 databases the visual information is represented

Author's personal copy

by model-based features, while image-based features are employed in AVLetters database. In addition, the proposed classification scheme is extended for the case of considering multiple and simultaneous streams of speech data. In particular, it is evaluated by considering three synchronized streams composed by audio, visual and audio–visual information, respectively. Experimental results show that a good performance is achieved with the proposed system over the three databases and different types of input data. Additionally, the proposed method performs better than other reported methods in the literature over the two public databases. All the experiments were performed using the same configuration parameters. It is important to note that, in addition to the good performance achieved, the proposed method has the advantage of using the same configuration, avoiding the need for adapting the parameters in the wavelet-based representation stage or the ones in the RF classifier stage for each particular database.

The rest of this paper is organized as follows. The proposed speech classification scheme is described in Sect. 2. The databases employed to evaluated the proposed method are presented in Sect. 3. In Sect. 4, the evaluation protocol and experimental results are presented. Finally, some concluding remarks and perspectives for future study are included in Sect. 5.

2 Proposed classification front-end

As schematically depicted in Fig. 1, the proposed classification scheme consists of two main stages, namely, the wavelet-based representation and Random Forests classification blocks, respectively. Speech is a time varying signal, where utterances, even of the same word, have different temporal durations. On the other hand, Random Forests classifiers require fixed-length input data. Thus, audio–visual speech information must be represented with a fixed-length structure.

In the first stage, the time varying input parameters are resampled and then a multilevel decomposition is computed. This decomposition is performed based on Discrete Wavelet Transform. The idea is to obtain a compact representation of the input parameters. This is done by representing each time varying input parameter with their associated approximation coefficients resulting from the DWT. In this way, independently of the number of frames associated with each word, a resulting fixed length feature vector is obtained. This method is also independent of the kind and length of the input data. In this paper, it is evaluated considering acoustic, visual and fused audio–visual input information separately. The length of the resulting feature vector is related to the chosen resolution level, which obviously determines the approximation accuracy. Since the length of the feature vector has to be kept reasonably small, there will be a trade-off between accuracy and feature vector length. For the DWT, the widely used db4 wavelet (Daubechies 1992) is employed. In Fig. 2, a signal and the corresponding approximations using db4 wavelet with decomposition levels $l = 2$ and $l = 4$ are shown. In Fig. 3, the wavelet-based feature vector computation method is illustrated, for the case of considering speech information represented by 4 time varying input parameters (or signals), denoted as $c_1(t)$, $c_2(t)$, $c_3(t)$, and $c_4(t)$, respectively. The wavelet-based feature vector is generated in two steps. First, each input parameter is resampled and its associated wavelet-based approximation coefficients, here denoted as $\overline{\mathbf{c}}_{1w}$, $\overline{\mathbf{c}}_{2w}$, $\overline{\mathbf{c}}_{3w}$ and $\overline{\mathbf{c}}_{4w}$, are computed. Then, these vectors are combined to form the resulting feature vector that is used as input to the Random Forests classifier.

Random Forests Breiman (2001) is an ensemble of decision trees. Since decision trees are very unstable, generally a small change in the dataset results in large changes in the developed model (Breiman 1996). The ensemble construction strategy is focused on increasing the diversity among the trees. The diversity is increased by fitting each tree on a bootstrap replicate (random subset of the available data, of the same length, taken with replacement) of the whole data. In addition, more diversity is introduced during the growing of each tree. The method selects a small random subset of P attributes for each node, and uses only this subset to search for the best split. The combination of these two sources of diversity produces an ensemble with good prediction performance. The performance will depend on the correlation between any two trees in the forest and on the strength of each individual tree. The stronger the individual trees are and the less correlated they are, the better error rate the classifier will achieve. The main parameters to adjust for

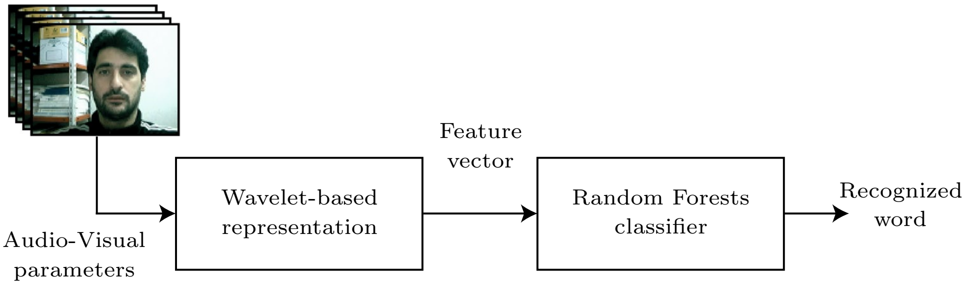


Fig. 1 Schematic representation of the proposed audio–visual speech classification system

Author's personal copy

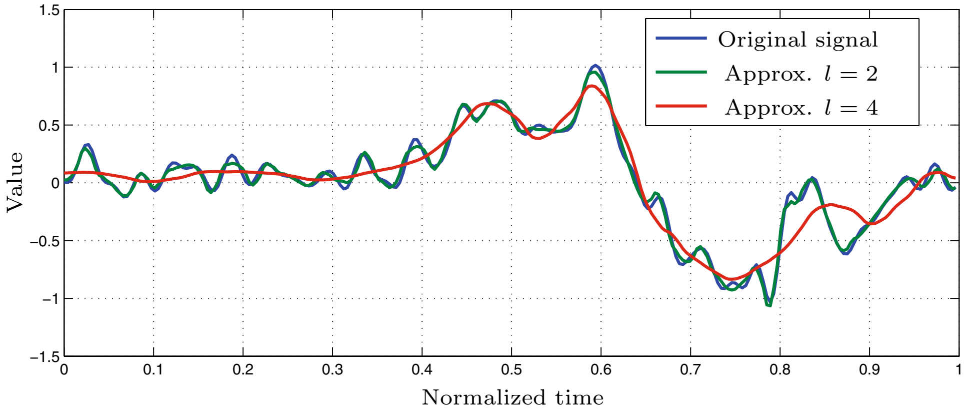


Fig. 2 Example of signal approximation using db4 wavelet with levels $l = 2$ and $l = 4$. (Color figure online)

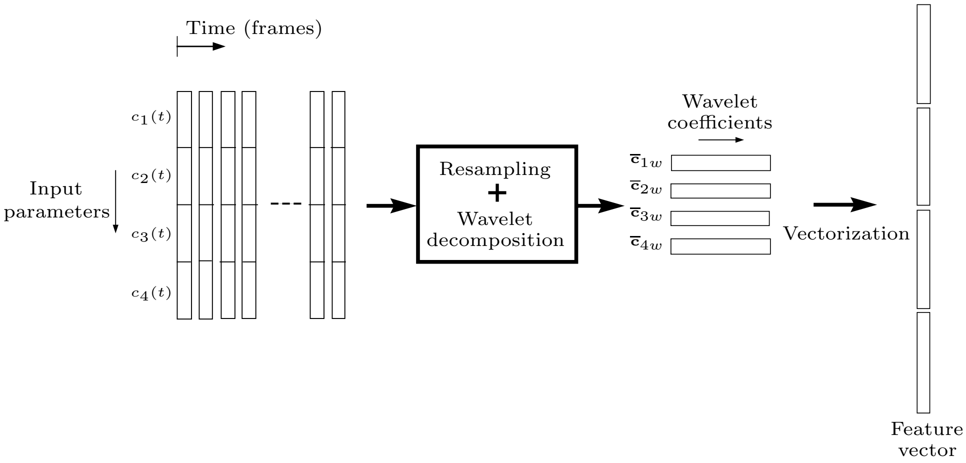


Fig. 3 Schematic representation of the method proposed for computing the wavelet-based feature vector. In this example, the input information is composed by 4 time varying parameters

a Random Forests classifier are the number of trees to grow and the number of randomly selected splitting variables to be considered at each node. The number of trees to grow does not strongly influence the results as long as it is kept large (generally, 2000 trees are enough). Then, in practice, the only tuning parameter of the RF classifier is the number of randomly selected splitting variables to be considered at each node, hereafter denoted as α.

In Sect. 4, this classification scheme is evaluated over different databases and considering different kinds of input speech data. As described bellow, for all the experiments, this evaluation is performed using fixed tuning parameters for the wavelet-based representation block, and the influence of parameter α is also analyzed.

3 Audio–visual databases

The evaluation of the proposed classification scheme is performed over three different audio–visual databases. One database was compiled by the authors of this paper and the remaining two are well known public databases. In particular, visual information during speech is extracted using an image-based method for one of the databases, while model-based visual features are considered for the remaining two. In the following, these databases are described.

3.1 AV-CMU database

The AV-CMU database (Huang and Chen 1998) consists in the recording of a series of words, uttered by ten speakers. In this paper, a subset of ten words (numbers from 1 to 10) is considered for the experiments. Each person spoke each word ten times, resulting in a total of 1000 utterances. The raw audio data is in the form of pulse-code-modulation-coded signals sampled at 44.1 kHz. The visual data is composed of the horizontal and vertical positions of the left (x_1, y_1) and right (x_2, y_2) corners of the mouth, as well as of the heights of the openings of the upper $(h1)$ and lower $(h2)$ lips, as depicted in Fig. 4a. The visual information was captured with a sample rate of 30 frames per seconds. To represent the visual information, the weighted least-squares parabolic fitting method proposed in Borgström and Alwan (2008) is employed in this paper. Visual features are then represented by five parameters, viz., the focal parameters

Author's personal copy

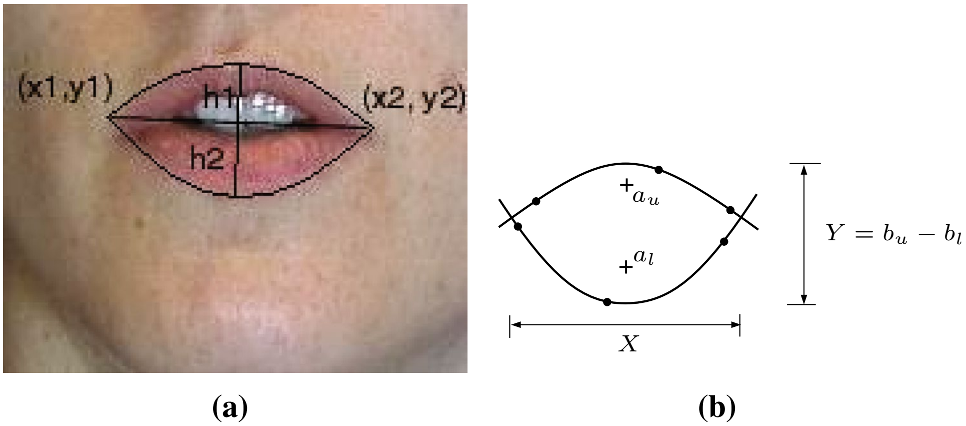


Fig. 4 CMU database. **a** Visual data included in the database. **b** Parabolic lip contour model proposed in Borgström and Alwan (2008)

of the upper and lower parabolas, mouths width and height, and the main angle of the bounding rectangle of the mouth.

3.2 AV-1 database

The AV-1 database consists of videos of 16 speakers, pronouncing a set of ten words in random order. These words correspond to the utterances of the following actions: *up*, *down*, *right*, *left*, *forward*, *back*, *stop*, *save*, *open* and *close*. Each word was pronounced 20 times by each speaker, resulting in a total of 3200 utterances. This database was compiled by the authors of this paper. The videos were recorded at a rate of 60 frames per second with a resolution of 640 × 480 pixels, and the audio was recorded at 8 kHz synchronized with the video. Visual features are extracted using the method proposed in (Terissi and Gómez 2010), which is based on a simple 3D face model, namely *Candide-3* (Ahlberg 2001), widely used in computer graphics, computer vision and model-based image-coding applications. For each frame of the videos, 3 parameters are computed and used to represent the visual information, viz., mouth height (v_H), mouth width (v_W) and area between lips (v_A), as depicted in Fig. 5.

3.3 AVLetters database

The AVLetters database Matthews et al. (2002) consists of three repetitions by each of ten speakers, five males (two with moustaches) and five females, of the isolated letters A–Z, resulting in a total of 780 utterances. This database provides pre-extracted mouth region of 80 × 60 pixels. It does not provide the original acoustic voice signals, but it includes the Mel-Frequency Cepstrum Coefficients (MFCC) associated to each uttered word. Figure 6 shows example images from the ten speakers. Visual information associated with speech is extracted and represented using the method based on local spatiotemporal descriptors proposed in Zhao et al. (2009). As described in Zhao et al. (2009), spatiotemporal local binary patterns are extracted from mouth regions and then used for describing words being uttered, taking into account the motion of mouth region and time order in pronunciation. 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.

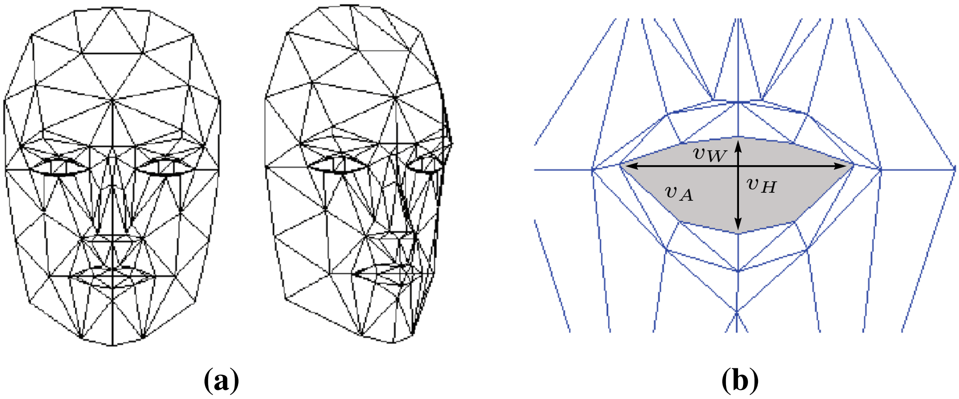


Fig. 5 AV-1 database. **a** *Candide-3* face model. **b** Visual parameters

Author's personal copy

Fig. 6 AVLetters database. Example images of the ten speakers

4 Experimental results

As mentioned before, the proposed speech recognition scheme can handle different types of input data. Hence, the performance of the proposed scheme is evaluated separately in different scenarios, viz., considering only audio information, only video information (lip-reading), and fused audio–visual information, respectively. These evaluations are carried out over the three databases described in the previous section, by computing the word recognition rates. The results shown in this section correspond to small vocabularies (between 10 and 24 words), but the proposed approach can be employed for medium and large vocabularies as well. In addition, the proposed recognition scheme is extended for the case of considering multiple and simultaneous streams of speech data. In particular, it is evaluated considering three synchronized streams composed by audio, visual and audio–visual information, respectively.

4.1 Evaluation protocol

Independently of the scenario and database being considered, the tuning parameters of the system are the ones associated with the wavelet-based feature representation block and the ones corresponding to the RF classifier. Regarding the wavelet-based representation, the tuning parameters are the normalized length of the resampled time functions, the mother wavelet and the resolution level for the approximation. In all the experiments presented in this paper, these parameters remained fixed. In particular, the normalized length was set to 256, the wavelet resolution level was set to 3, and the widely used db4 was chosen as the mother wavelet. Considering these values, each input parameter is represented by its associated 38 wavelet-based approximation coefficients. Regarding the RF classifier, the parameters to adjust are the number of trees to grow and the number of randomly selected splitting variables to be considered at each node. However, the number of trees to grow does not strongly influence the performance of the classifier as long as it is kept large. In particular, in the experiments presented in this paper this value is set to 2000 trees. Thus, the only tuning parameter of the proposed recognition scheme is the number of randomly selected splitting variables to be considered at each node, denoted as α. In all the experiments, the proposed scheme is evaluated with values of α in the range from 2 to 10.

In order to obtain statistically significant results, at each experiment a D-fold cross-validation (CV) is performed over the whole data to compute the recognition rates. For the cases of AV-CMU and AV-1 databases, at each fold, one speaker is used for testing and the remaining ones for training, resulting in a speaker independent evaluation. Thus, a 10-fold CV is used for the AV-CMU database (10 speakers) and a 16-fold CV for the AV-1 database (16 speakers).

For the case of the AVLetter database, the evaluation is performed with the same protocol employed in other approaches reported in the literature (Matthews et al. 2002; Zhao et al. 2009; Amer et al. 2014; Hu et al. 2016) over this database. In this case, the training set is composed by the first two utterances of each letter spoken by each speaker, and the remaining ones are used for the testing set. Hence, the training set and testing set both contain the same set of speakers, resulting in a speaker dependent evaluation.

For comparison purposes, these experiments are also performed with other three classification approaches, viz., traditional Hidden Markov Models (Rabiner 1989) (HMMs), Support Vector Machines (Cortes and Vapnik 1995) (SVM) and Boosting-based classification. These approaches are briefly described in the following subsections.

4.1.1 HMM

Since Rabiner proposed the HMMs for speech recognition (Rabiner 1989), these statistical generative models have become the traditional framework in automatic speech recognition applications. HMMs were specifically designed to capture the evolution of the temporal dynamics in the observed data. For that reason, the acoustic and visual

Author's personal copy

features extracted from speech can be used directly as inputs to the HMMs. For comparison purposes, in this work the HMMs are implemented using N-state left-to-right models and considering continuous symbol observation, represented by the linear combination of M Gaussian distributions. At each experiment, the parameters associated with the HMMs are optimized via exhaustive search, considering N in the range from 1 to 20, and M from 1 to 30, looking for the combination of number of states (N) and number of Gaussian combinations (M) leading to the best performance.

4.1.2 SVM

Support Vector Machines (Cortes and Vapnik 1995) are supervised discriminative learning models that have been used in many classification applications. SVMs define the best separating hyperplane by maximizing the margin between boundary points of the classes and the separating hyperplane. These boundary points are referred as support vectors. SVMs use linear and nonlinear separating hyperplanes for data classification. In this work, SVMs are implemented by considering Gaussian kernels. Thus, the tuning parameters are the cost C and the σ value of the Gaussian kernel. Similarly to the case of Random Forests classifiers, SVMs require fixed length input data. For that reason, the same procedure based on Discrete Wavelet Transform described in Sect. 2, is used to obtain audio–visual speech features with a fixed-length structure. Hereafter, these classification models will referred as W+SVM. At each experiment, the tuning parameters associated with W+SVMs are optimized via exhaustive search. The search of optimum values for C and σ was carried out in two stages. First, a search is carried out varying C and σ parameters values with decade steps ($[\dots, 10^2, 10^1, 1, 10^2, \dots]$). Then, in the region where the best results are found, a second finest search with smaller step is carried out to find the final optimized values for C and σ.

4.1.3 AdaBoost

Adaptive Boosting is a machine learning algorithm proposed in (Schapire and Singer 1999) based on the idea of creating a highly accurate prediction rule by combining many relatively weak and inaccurate rules. Usually, decision trees are used as weak classifiers. AdaBoost also requires fixed length input data, thus in this paper, the same procedure based on DWT employed for the cases of RF and SVM, is used to obtain audio–visual speech features with a fixed-length structure. Hereafter, these classification models will referred as W+ADA. The parameters to adjust in AdaBoost classification are the number of iterations N of the boosting algorithm and the depth of each tree in the ensemble d. At each experiment presented in this paper, the parameters are optimized via exhaustive search, considering $N = [100, 500, 1000, 2000, 5000]$ and d in the range from 2 to 10, looking for the combination leading to the best performance.

4.2 Acoustic noisy conditions

The presence of noise in the acoustic signal is the main source of variability affecting the performance of speech recognition systems. It is important then to analyze the recognition systems in the presence of noise in the acoustic channel. For that reason, in all the experiments where acoustic information is considered, that only excludes the experiments corresponding to lip-reading, the proposed classification scheme is evaluated by considering noisy acoustic conditions. To do so, experiments with additive Gaussian and additive Babble noise, with signal-to-noise ratios (SNRs) ranging from − 10 to 40 dB, are performed. Multispeaker or Babble noise environment is one of the most challenging noise conditions, since the interference is speech from other speakers. This noise is uniquely challenging because of its highly time evolving structure and its similarity to the desired target speech (Krishnamurthy and Hansen 2009). In this paper, Babble noise samples were extracted from *NOISEX-92* database, compiled by the Digital Signal Processing (DSP) group at Rice University (Varga and Steeneken 1993). In these experiments, the system is trained using clean audio information and then it is evaluated with acoustic signals with different SNRs.

4.3 Visual information

In this subsection, the performance of the proposed recognition scheme for the case of considering only visual speech information is analyzed. This evaluation is performed separately over the three databases described in Sect. 3.

4.3.1 AV-CMU database

In Table 1, the recognition rates obtained over a subset of ten words (numbers from 1 to 10), with proposed recognition scheme are presented. The two values correspond to the minimum and maximum performances of the proposed system for the cases of setting the only tuning parameter α, in the range from 2 to 10. This table also includes the recognition rates obtained with traditional HMMs, and the ones obtained by considering SVM and Boosting as classification method. Additionally, the performance reported in Borgström and Alwan (2008) over the same database is included. In all cases, visual features are represented based on the parabolic contour lip model briefly described in Sect. 3.1. As can be observed, the proposed method performs better

Author's personal copy

than the other approaches independently of the value used for variable α.

4.3.2 AVLetters database

In Table 2, a comparison of the recognition results obtained with the method proposed in this paper (last row) with the corresponding ones obtained with other methods proposed in the literature, is shown. It can be observed that using the local spatiotemporal descriptors as visual features (rows 2, 3, 6, 7 and 8), the proposed W+RF approach yields better accuracy than the methods based on HMM and SVM reported in Zhao et al. (2009), and also performs better than the approaches based on W+SVM and W+ADA. In addition, the proposed approach also achieves better results in comparison to the method in Matthews et al. (2002) based on HMM classifier and multiscale spatial analysis, and the ones presented in Ngiam et al. (2011) and Hu et al. (2016) based on deep networks and Recurrent Temporal Multimodal Restricted Boltzmann Machine (RTMRBM), respectively.

4.3.3 AV-1 database

This database, where the visual information is represented by mouth shape parameters, was also employed to evaluate the performance of the proposed classification scheme for lip-reading. These results are presented in Table 3. This table also includes the performance obtained with traditional HMM, W+SVM and W+ADA classification schemes. These results also show that, for any value of α, the proposed W+RF scheme performs better than the other methods.

Table 1 Lip-reading on AV-CMU database

Classifier	Visual features	Accuracy (%)
HMM (Borgström and Alwan 2008)	Lip Contour Model (Borgström and Alwan 2008)	61.17
HMM	Lip Contour Model (Borgström and Alwan 2008)	57.79
W+ADA	Lip Contour Model (Borgström and Alwan 2008)	67.53
W+SVM	Lip Contour Model (Borgström and Alwan 2008)	67.53
Proposed W+RF	Lip Contour Model (Borgström and Alwan 2008)	**69.59–71.65**

Experiments performed over a subset of words, corresponding to the numbers from 1 to 10. For the proposed classification scheme, the values correspond to the minimum and maximum performances obtained with α in the range from 2 to 10

Table 2 Lip-reading on AVLetters database

Classifier	Visual features	Accuracy (%)
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
W+SVM	Local spatiotemporal descriptors (Zhao et al. 2009)	63.08
W+ADA	Local spatiotemporal descriptors (Zhao et al. 2009)	54.23
Proposed W+RF	Local spatiotemporal descriptors (Zhao et al. 2009)	**61.12–65.38**

For the proposed classification scheme, the values correspond to the minimum and maximum performances obtained with α in the range from 2 to 10

Table 3 Lip-reading on AV-1 database

Classifier	Visual features	Accuracy (%)
HMM	Mouth shape parameters (Terissi and Gómez 2010)	70.16
W+ADA	Mouth shape parameters (Terissi and Gómez 2010)	85.63
W+SVM	Mouth shape parameters (Terissi and Gómez 2010)	83.75
Proposed W+RF	Mouth shape parameters (Terissi and Gómez 2010)	**85.21–88.67**

For the proposed classification scheme, the values correspond to the minimum and maximum performances obtained with α in the range from 2 to 10

Author's personal copy

The above presented results show that, over the three databases, the proposed recognition scheme performs satisfactorily and better than other approaches reported in the literature. It is also important to note that the efficiency of the proposed scheme is slightly affected by using different values for the only tuning parameter of the system. This indicates that, in contrast to other methods described in the literature, there is no need to perform an optimization of the system for each particular dataset. The results for the HMM approach are the worst for the three databases compared with the other methods. This suggests that HMM are not appropriated to represent visual information during speech, in word recognition applications.

4.4 Acoustic information

The evaluation of the proposed classification scheme for the case of considering only acoustic information is presented in this section. As described in Sect. 4.2, this evaluation is carried out in noisy acoustic conditions, in particular, considering additive Gaussian and additive Babble noises. The AVLetters database does not include the original acoustic voice signals, but it includes the Mel-Cepstral coefficients associated to each uttered word without acoustic noise. Thus, it is not possible to perform the evaluation by considering acoustic noisy conditions. For this reason, the evaluation over AVLetters database is performed considering only clean acoustic information. For AV-CMU and AV-1 databases the audio signal is partitioned in frames. For each frame, the audio features are represented by the first eleven non-DC Mel-Cepstral coefficients, and their associated first and second time derivatives. Similarly to the case of considering only visual information, the experiments are performed considering different values for parameter α in the range from to 2 to 10. Additionally, these experiments are carried out with traditional HMMs, W+SVM and W+ADA classification schemes, in order to compare their performances with the proposed one.

In Table 4, the performances obtained over the AVLetters database are shown. These results show that the proposed method outperforms the classification efficiency obtained with other approaches proposed in the literature. The recognition rates obtained at different SNRs over the AV-CMU and AV-1 databases are depicted in Fig. 7. In this figure, the performance of the proposed classification scheme is represented by the red area, which corresponds to the recognition rates obtained by selecting α in the range from 2 to 10. As expected, for all the cases, the performance in the recognition task deteriorates as the SNR decreases. This is due to the mismatch between training (noiseless) and testing (noisy) acoustic data.

Table 4 Classification over AVLetters database considering only acoustic information

Approach	Accuracy (%)
MDAE (Ngiam et al. 2011)	58.40
CRBM (Amer et al. 2014)	61.20
RTMRBM (Hu et al. 2016)	64.41
W+ADA	55.77
W+AVM	62.31
Proposed W+RF	**65.77–71.54**

For the proposed classification scheme, the values correspond to the minimum and maximum performances obtained with α in the range from 2 to 10

It is clear that, for both databases and both noise types, the proposed recognition scheme performs better than the ones based on HMMs, W+SVM and W+ADA, mainly at low SNRs. These results also show that the value of parameter α does not modify significantly the performance of the proposed W+RF scheme. It can be observed also that for the case of clean audio signal, all the methods work properly, but the HMM-based ones perform slightly better.

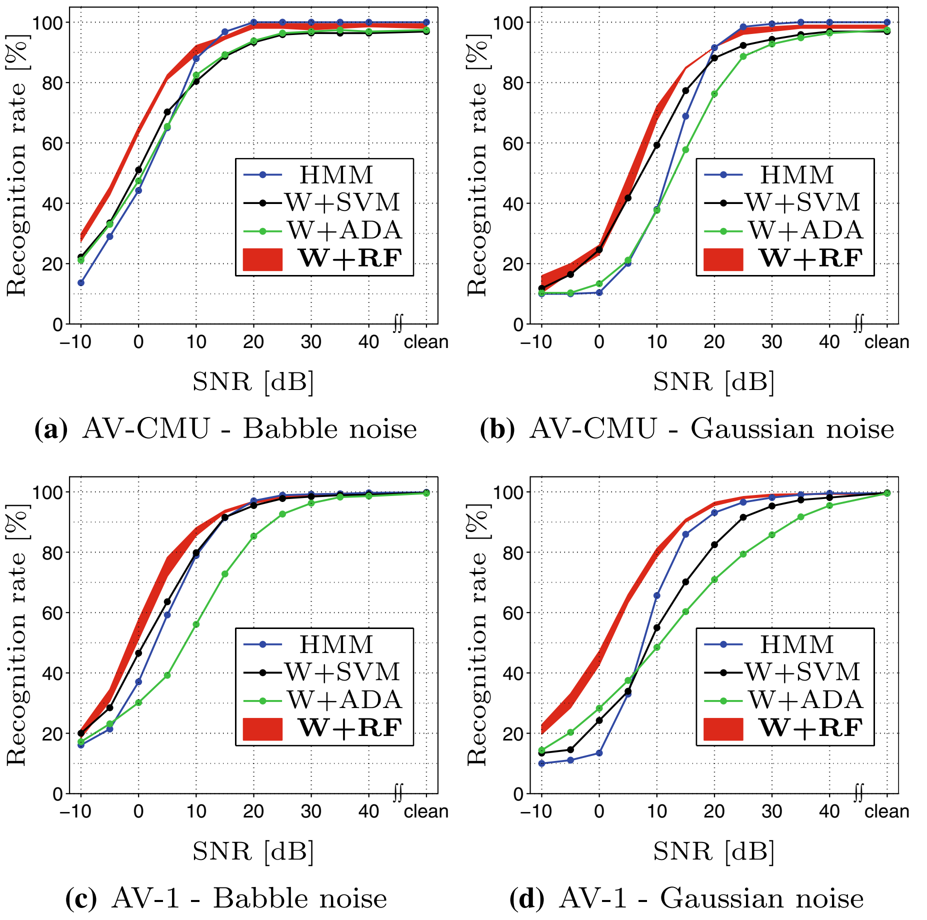
4.5 Fused audio–visual information

In this subsection, the evaluation of the proposed scheme for the case of considering speech represented by fused audio–visual information is presented. Similarly to the case of considering only audio information, the experiments are performed in noisy acoustic conditions over AV-CMU and AV-1 databases, and at clean acoustic conditions for the AVLetters database. The audio signal is partitioned in frames with the same rate as the video frame rate. Then, the fused audio–visual feature vector at frame t is composed by the concatenation of the corresponding acoustic and visual parameters at frame t. Once again, the evaluation of the proposed classification scheme is performed considering different values for parameter α in the range from 2 to 10, and these experiments are also carried out with traditional HMMs, W+SVM and W+ADA classification schemes, in order to compare their performance with the proposed one.

The classification performances obtained over the AVLetters database are shown in Table 5. These results show that the proposed W+RF approach performs better in comparison with other methods proposed in the literature, independently of the value used for parameter α. On the other hand, the recognition rates obtained at different SNRs over the AV-CMU and AV-1 databases are depicted in Fig. 8. The results show that, for both databases and both noise types, the performance of the proposed classification scheme is better than the ones corresponding to methods based on HMMs, W+SVM and W+ADA. Performance improvements are more notorious at low and middle range of SNRs. It is also clear from these results that using any value of parameter α leads to good performances of the proposed system.

Author's personal copy

Fig. 7 Classification based on acoustic information. Recognition rates obtained over the AV-CMU (first row) and AV-1 (second row) databases for different SNRs for the cases of considering Gaussian and Babble noises. The performance of the proposed classification scheme (W+RF) is represented by the red area, which corresponds to the recognition rates obtained by selecting α in the range from 2 to 10. (Color figure online)



(a) AV-CMU - Babble noise

(b) AV-CMU - Gaussian noise

(c) AV-1 - Babble noise

(d) AV-1 - Gaussian noise

Comparing the results in Figs. 7 and 8, it can be noted that the use of audio–visual information improves the recognition rates for all the classification approaches.

4.6 Multi-stream information

The proposed classification scheme can be also employed in scenarios where multiple streams or modalities of speech information are available. For these cases, a simple parallel configuration is presented in this paper, where independent classifiers are used for each modality and the final decision is computed by the combination of the likelihood scores associated with each modality. In the literature, this is known as late integration or decision level fusion (Lee and Park 2008; Estellers et al. 2012). This strategy does not require strictly synchronized streams. Different techniques to perform decision level fusion have been proposed. The most commonly used is to combine the matching scores of the individual classifiers with simple rules, such as, *max*, *min*, *product*, or *weighted sum*.

Table 5 Classification over AVLetters database considering fused audio–visual information

Approach	Accuracy (%)
MDAE (Ngiam et al. 2011)	62.90
CRBM (Amer et al. 2014)	64.80
RTMRBM (Hu et al. 2016)	66.04
W+ADA	60.34
W+SVM	68.22
Proposed W+RF	**68.85–72.69**

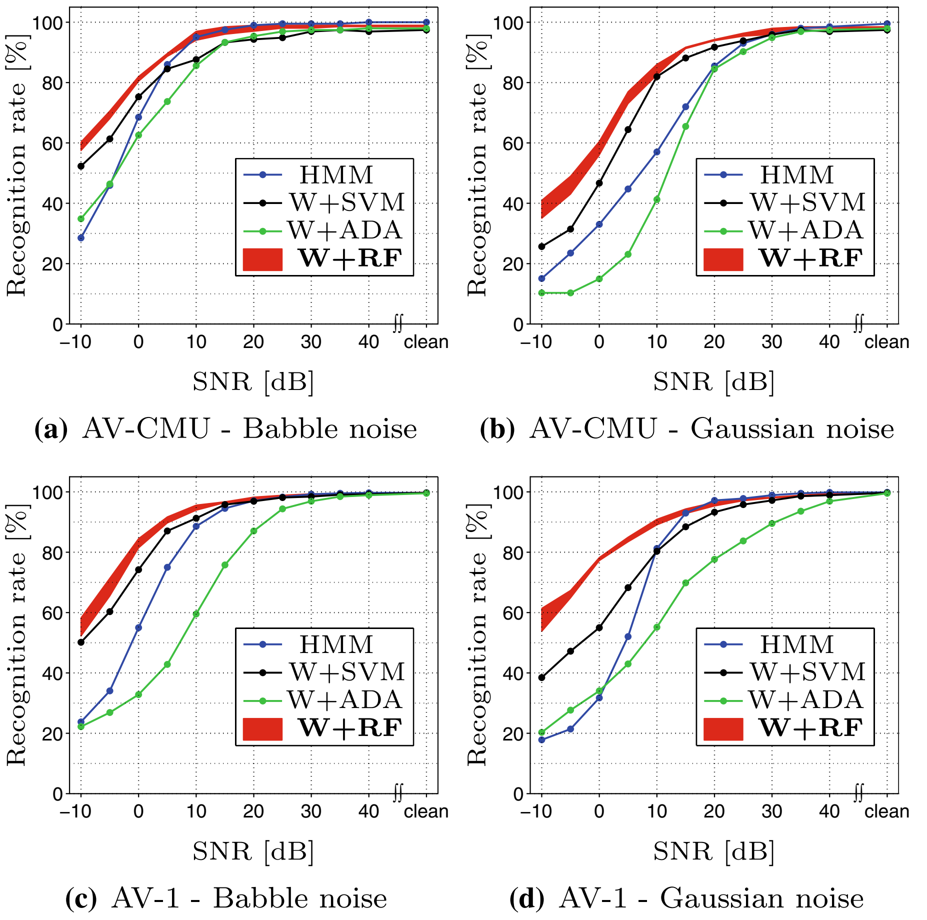
For the proposed classification scheme, the values correspond to the minimum and maximum performances obtained with α in the range from 2 to 10

In Fig. 9, a schematic representation of the proposed configuration for handling N input streams is depicted. Considering N input observations associated to different modalities, denoted as O_1, O_2, ..., O_N, respectively, N independent W+RF classifiers are employed, denoted as λ_1, λ_2, ..., λ_N, respectively. Given a set of observations associated to a word to be recognized, the probability (or score) vectors $\mathbf{P}\left(O_1|\lambda^1\right)$, $\mathbf{P}\left(O_2|\lambda^2\right)$, ..., $\mathbf{P}\left(O_N|\lambda^N\right)$ are computed for each modality. These vectors are composed by the concatenation of the probabilities associated with each class in the dictionary. Then, the fused probability vector $\mathbf{P}_F\left(O_1, O_2, \dots, O_N\right)$ is computed as

$$\mathbf{P}_F\left(O_1, O_2, \dots, O_N\right) = \\ = \frac{1}{N}\left[\mathbf{P}\left(O_1|\lambda^1\right) + \mathbf{P}\left(O_2|\lambda^2\right) + \cdots + \mathbf{P}\left(O_N|\lambda^N\right)\right].$$

Author's personal copy

Fig. 8 Classification based on fused audio–visual information. Recognition rates obtained over the AV-CMU (first row) and AV-1 (second row) databases for different SNRs for the cases of considering Gaussian and Babble noises. The performance of the proposed classification scheme (W+RF) is represented by the red area, which corresponds to the recognition rates obtained by selecting α in the range from 2 to 10. (Color figure online)



(a) AV-CMU - Babble noise

(b) AV-CMU - Gaussian noise

(c) AV-1 - Babble noise

(d) AV-1 - Gaussian noise

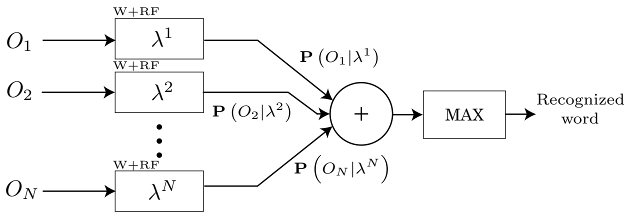

O_1
O_2
O_N
W+RF
W+RF
W+RF
λ^1
λ^2
λ^N
$\mathbf{P}(O_1|\lambda^1)$
$\mathbf{P}(O_2|\lambda^2)$
$\mathbf{P}(O_N|\lambda^N)$
+
MAX
Recognized word

Fig. 9 Schematic representation of the proposed configuration for handling multiple streams of audio and visual speech information

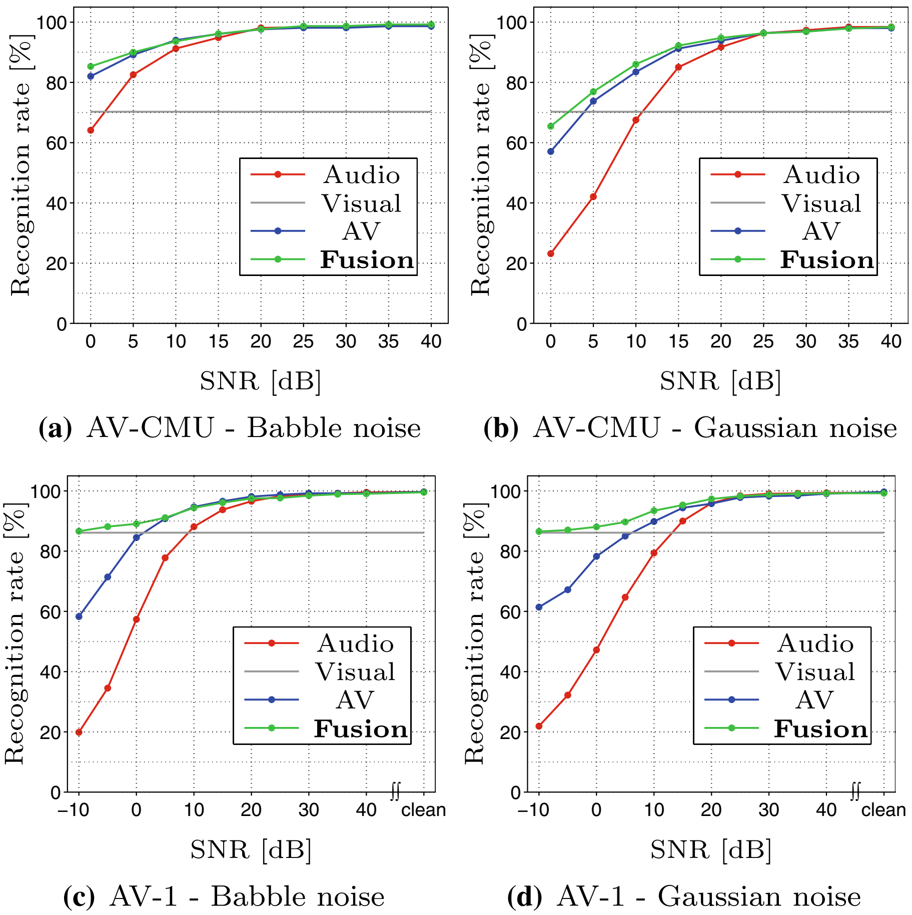
Finally, as illustrated in Fig. 9, the input data is recognized as the class with the maximum fused probability.

This multi-stream configuration is evaluated by considering acoustic, visual and fused audio–visual streams simultaneously. The evaluation is carried out over the AV-CMU and AV-1 databases. Again, the experiments are performed in noisy acoustic conditions, considering additive Gaussian and additive Babble noises. The results of this evaluation are shown in Fig. 10. As shown in the previous subsections, the value of parameter α does not affect significantly the performance of the proposed W+RF classifier scheme in the cases of considering audio, visual or audio–visual information. For clarity reasons, the results depicted in Fig. 10 correspond to the cases of considering parameter $\alpha = 4$, since the performances obtained for different values of α have a similar behaviour. As can be observed in Fig. 10a and b for the case of AV-CMU database, the multi-stream approach leads to a slightly enhancement of the recognition rates in comparison to the ones obtained by considering fused audio–visual information. However, for the case of AV-1 database, see Fig. 10c and d, the combination of the three streams enforces a significant improvement of the recognition rates in comparison with the ones obtained with individual classifiers.

The AV-CMU database was also employed in Borgström and Alwan (2008) to evaluate a multi-stream classification approach based on the combination of HMMs. Figure 11 compares the performances obtained with the method reported in Borgström and Alwan (2008) and the ones obtained with the proposed multi-stream configuration. It is clear that the proposed method outperforms the one in Borgström and Alwan (2008) across all the considered SNRs.

It must be noted that the performance of the proposed multi-stream classifier could be improved, for instance, by combining the individual probabilities taking into account the level of noise of the acoustic stream, and thus more importance could be given to the visual information in cases

Author's personal copy



(a) AV-CMU - Babble noise

(b) AV-CMU - Gaussian noise

(c) AV-1 - Babble noise

(d) AV-1 - Gaussian noise

Fig. 10 Classification based on multi-stream information. Recognition rates obtained over the AV-CMU (first row) and AV-1 (second row) databases for different SNRs for the cases of considering Gaussian and Babble noises. The performance of the proposed recognition multi-stream configuration is depicted in green line. Performances for the cases of considering individual W+RF classifiers based on audio, video and fused audio–visual information are depicted in red, grey and blue lines, respectively. (Color figure online)

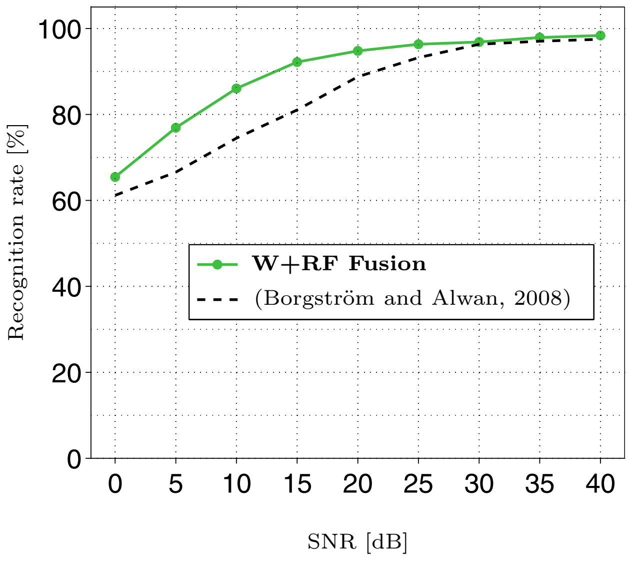


Fig. 11 Efficiency comparison over the AV-CMU database between the proposed multi-stream approach and the one presented in Borgström and Alwan (2008)

of low SNR. However, this kind optimization is out of the scope of this paper. These experiments are included in order to show that the proposed W+RF classification method can be also employed in a multi-stream scenario satisfactorily.

4.7 Overall analysis

The above presented results show that the proposed classification scheme performs satisfactorily for classifying speech regarding to different speech features and databases. In addition, it yields better recognition rates than other methods proposed in the literature. However, the most prominent characteristic of the proposed method is that it can be employed in different scenarios, without requiring an optimization of its meta-parameters. As stated before, all the experiments presented in this paper have been carried out using fixed configuration parameters, except for the case of parameter α, which value does not affect significantly the recognition performance. It must be noted that this is not the usual situation of other approaches reported in the literature. On the contrary, the performance of other approaches usually depends on specific configurations for a particular database, either based on the number and size of the input features, the number of words being considered, the amount

Author's personal copy

of examples to train the system, the inter and intra class variability of the data, etc. In this sense, the proposed classification scheme was evaluated considering different conditions. For instance, for the case of considering only visual information (lip reading) over the AV-1 database, each word is characterized by three visual temporal parameters, resulting (after the wavelet-based stage) in a fixed-length feature vector composed by 114 coefficients, while for the case of considering fused audio–visual information over the AV-CMU dabatase, where each word is characterized by 38 parameters (5 visual + 33 acoustic), the feature vectors are composed by 1444 coefficients. On the other hand, experiments on AV-CMU and AV-1 databases were carried out considering 10 classes, while 26 were used for the AVLetters database.

5 Conclusions

In this paper, a speech classification front-end based on wavelets and Random Forests has been proposed. This system can be employed for recognizing speech using acoustic, visual or audio–visual information. Wavelet multiresolution analysis is used to represent in a compact form the sequence of acoustic and visual input data. The wavelet-based features are employed as inputs of a Random Forest classifier to perform the speech recognition. 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 by the authors of this paper. Experimental results show that a good performance is achieved with the proposed system over the three databases. In addition, the proposed method performs better than other reported methods in the literature over the same two public databases. All the experiments presented in this paper have been carried out using fixed configuration parameters, except for the case of parameter α, for which values in the range from 2 to 10 has been considered. Experimental results show that selecting α in this range does not affect significantly the recognition performance. Thus, there was no need to adapt the wavelet decomposition parameters or the Random Forests classifier parameters to each particular database or experiment. This is an important advantage of the proposed approach in comparison to other methods that require an optimization stage of the classifiers meta-parameters.

Funding The funding was provided by the Agencia Nacional de Promoción Científica y Tecnológica Grant No. (PICT 2014-2041), Ministerio de Ciencia, Tecnología e Innovación Productiva Grant No. (STIC-AmSud Project 15STIC-05) and Universidad Nacional de Rosario Grant No. (Project Ing395).

References

Ahlberg, J. (2001). Candide-3: An updated parameterised face. Technical report, Linkoping: Department of Electrical Engineering, Linkping University.

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.

Aleksic, P., Williams, J., Wu, Z., & Katsaggelos, A. (2002). Audio-visual continuous speech recognition using MPEG-4 compliant visual features. In *Proceedings of the International Conference on Image Processing*, vol 1, pp. 960–963.

Ali, H., Ahmad, N., Zhou, X., Iqbal, K., & Ali, S. M. (2014). Dwt features performance analysis for automatic speech recognition of Urdu. *SpringerPlus*, *3*(1), 204.

Ali, H., Jianwei, A., & Iqbal, K. (2015). Automatic speech recognition of urdu digits with optimal classification approach. *International Journal of Computer Applications*, *118*(9), 1–5.

Amer, M. R., Siddiquie, B., Khan, S., Divakaran, A., & Sawhney, H. (2014). Multimodal fusion using dynamic hybrid models. In *IEEE Winter Conference on Applications of Computer Vision*, pp. 556–563.

Attar, M., Mosleh, M., & Ansari-Asl, K. (2010). Isolated words-recognition based on random forest classifiers. In *Proceedings of 2010 4th International Conference on Intelligent Information Technology*.

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.

Borde, P., Varpe, A., Manza, R., & Yannawar, P. (2015). Recognition of isolated words using Zernike and MFCC features for audio visual speech recognition. *International Journal of Speech Technology*, *18*(2), 167–175.

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. (1996). Bagging predictors. *Machine Learning*, *26*(2), 123–140.

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

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

Daubechies, I. (1992). *Ten Lectures on Wavelets*. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics.

Dong, L., Foo, S. W., & Lian, Y. (2005). A two-channel training algorithm for hidden Markov model and its application to lip reading. *EURASIP Journal on Advances in Signal Processing*, *2005*(9), 347367.

Dupont, S., & Luettin, J. (2000). Audio-visual speech modeling for continuous speech recognition. *IEEE Transactions on Multimedia*, *2*(3), 141–151.

Estellers, V., Gurban, M., & Thiran, J. (2012). On dynamic stream weighting for audio-visual speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, *20*(4), 1145–1157.

Farooq, O., & Datta, S. (2003a). Phoneme recognition using wavelet based features. *Information Sciences*, *150*(1–2), 5–15.

Farooq, O., & Datta, S. (2003b). Wavelet-based denoising for robust feature extraction for speech recognition. *Electronics Letters*, *39*(1), 163–165.

Foo, S., 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.

Author's personal copy

Gowdy, J., Subramanya, A., Bartels, C., & Bilmes, J. (2004). DBN based multi-stream models for audio-visual speech recognition. *In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, *1*, 993–996.

Gowdy, J. N. & Tufekci, Z. (2000). Mel-scaled discrete wavelet coefficients for speech recognition. In *IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.00CH37100)*, vol 3, pp. 1351–1354.

Gupta, M. & Gilbert, A. (2001). Robust speech recognition using wavelet coefficient features. In *IEEE Workshop on Automatic Speech Recognition and Understanding, 2001. ASRU '01.*, pp. 445–448.

Hu, D., Li, X., & Lu, X. (2016). Temporal multimodal learning in audiovisual speech recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3574–3582.

Huang, F. J. & Chen, T. (1998). Advanced Multimedia Processing Laboratory. Cornell University, Ithaca, NY. Accessed March 2018, from http://chenlab.ece.cornell.edu/projects/AudioVisualSpeechProcessing.

Iwano, K., Yoshinaga, T., Tamura, S., & Furui, S. (2007). Audio-visual speech recognition using lip information extracted from side-face images. *EURASIP Journal on Audio, Speech, and Music Processing*, *2007*(1), 064506.

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

Kotnik, B., Kacic, Z., & Horvat, B. (2003). The usage of wavelet packet transformation in automatic noisy speech recognition systems. In *The IEEE Region 8 EUROCON 2003. Computer as a Tool.*, vol. 2, pp. 131–134.

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

Lee, J.-S., & Park, C.-H. (2008). Robust audio-visual speech recognition based on late integration. *IEEE Transactions on Multimedia*, *10*(5), 767–779.

Maganti, H. K., & Matassoni, M. (2014). Auditory processing-based features for improving speech recognition in adverse acoustic conditions. *EURASIP Journal on Audio, Speech, and Music Processing*, *2014*(1), 21.

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*, 2002.

Miki, M., Kitaoka, N., Miyajima, C., Nishino, T., & Takeda, K. (2014). Improvement of multimodal gesture and speech recognition performance using time intervals between gestures and accompanying speech. *EURASIP Journal on Audio, Speech, and Music Processing*, *2014*(1), 2.

Monaci, G., Vandergheynst, P., & Sommer, F. T. (2009). Learning bimodal structure in audio-visual data. *IEEE Transactions on Neural Networks*, *20*(12), 1898–1910.

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 (ICML-11)*, pp. 689–696.

Panda, S. P., & Nayak, A. K. (2016). Automatic speech segmentation in syllable centric speech recognition system. *International Journal of Speech Technology*, *19*(1), 9–18.

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

Pavez, E., & Silva, J. F. (2012). Analysis and design of wavelet-packet cepstral coefficients for automatic speech recognition. *Speech Communication*, *54*(6), 814–835.

Petridis, S. & Pantic, M. (2016). Deep complementary bottleneck features for visual speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2304–2308.

Potamianos, G., Graf, H. P., & Cosatto, E. (1998). An image transform approach for HMM based automatic lipreading. In *Proceedings of the International Conference on Image Processing*, pp. 173–177.

Potamianos, G., Neti, C., Gravier, G., & Garg, A. (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.

Puviarasan, N., & Palanivel, S. (2011). Lip reading of hearing impaired persons using HMM. *Expert Systems with Applications*, *38*(4), 4477–4481.

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

Rajeswari, P. N. N. S. S., & Sathyanarayana, V. (2014). Robust speech recognition using wavelet domain front end and hidden Markov models. In V. Sridhar, H. S. Sheshadri, & M. C. Padma (Eds.), *Emerging research in electronics, computer science and technology*. New Delhi: Springer.

Saitoh, T., Morishita, K., & Konishi, R. (2008). Analysis of efficient lip reading method for various languages. In *Proceedings of the 19th International Conference on Pattern Recognition*, pp. 1–4.

Schapire, R. E., & Singer, Y. (1999). Improved boosting algorithms using confidence-rated predictions. *Machine Learning*, *37*, 80–91.

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.

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.

Trottier, L., Giguère, P., & Chaib-draa, B. (2015). Feature selection for robust automatic speech recognition: a temporal offset approach. *International Journal of Speech Technology*, *18*(3), 395–404.

Tufekci, Z., Gowdy, J. N., Gurbuz, S., & Patterson, E. (2006). Applied mel-frequency discrete wavelet coefficients and parallel model compensation for noise-robust speech recognition. *Speech Communication*, *48*(10), 1294–1307.

Uluskan, S., Sangwan, A., & Hansen, J. H. L. (2017). Phoneme class based feature adaptation for mismatch acoustic modeling and recognition of distant noisy speech. *International Journal of Speech Technology*, *20*, 799–811.

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.

Author's personal copy

Wang, S. L., Lau, W. H., & Leung, S. H. (2004). Automatic lip contour extraction from color images. *Pattern Recognition*, *37*(12), 2375–2387.

Wright, J., Ma, Y., Mairal, J., Sapiro, G., Huang, T. S., & Yan, S. (2010). Sparse representation for computer vision and pattern recognition. *Proceedings of the IEEE*, *98*(6), 1031–1044.

Yau, W. C., Kumar, D. K., & Arjunan, S. P. (2007). Visual recognition of speech consonants using facial movement features. *Integrated Computer-Aided Engineering-Informatics in Control, Automation and Robotics*, *14*(1), 49–61.

Yin, S., Liu, C., Zhang, Z., Lin, Y., Wang, D., Tejedor, J., et al. (2015). Noisy training for deep neural networks in speech recognition. *EURASIP Journal on Audio, Speech, and Music Processing*, *2015*(1), 2.

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