# Noisy Speech Recognition Based on Combined Audio-Visual Classifiers

Lucas D. Terissi(✉), Gonzalo D. Sad, Juan C. Gómez, and Marianela Parodi

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

**Abstract.** An isolated word speech recognition system based on audio-visual features is proposed in this paper. To enhance the recognition over different noisy conditions, this system combines three classifiers based on audio, visual and audio-visual information, respectively. The performance of the proposed recognition system is evaluated over two isolated word audio-visual databases, a public one and a database compiled by the authors of this paper. Experimental results show that the structure of the proposed system leads to significant improvements of the recognition rates through a wide range of signal-to-noise ratios.

**Keywords:** Speech recognition · Audio-visual speech features · Audio-visual information fusion

## 1 Introduction

The last decades have witnessed an increasing interest in the development of more natural Human Computer Interfaces (HCI), that mimic the way humans communicate among themselves. Communication among humans is inherently a multimodal process, in the sense that, for the transmission of an idea, not only is important the acoustic signal but also the facial expressions and body gestures [6]. For instance, a significant role in spoken language communication is played by lip reading. This is essential for the hearing-impaired people, and is also important for normal listeners in noisy environments to improve the intelligibility of the speech signal. Audio Visual Speech Recognition (AVSR) is a fundamental task in HCIs, where the acoustic and visual information (mouth movements, facial gestures, etc.) during speech are taken into account. Several strategies have been proposed in the literature for AVSR [7–9], where improvements of the recognition rates are achieved by fusing audio and visual features related to speech. As expected, these improvements are more notorious when the audio channel is corrupted by noise, which is a usual situation in speech recognition applications. These approaches are usually classified according to the method employed to combine (or fuse) the audio and visual information. Three main approaches can be distinguished, *viz.*, feature level fusion, classifier


F. Schwenker et al. (Eds.): MPRSS 2014, LNAI 8869, pp. 43–53, 2015.
DOI: 10.1007/978-3-319-14899-1_5

level fusion and decision level fusion [4]. In feature level fusion (early integration), audio and visual features are combined to form a unique audio-visual feature vector, which is then employed for the classification task. This strategy requires the audio and visual features to be exactly at the same rate and in synchrony, and it is effective when the combined modalities are correlated, since it can exploit the covariations between audio and visual features. In classifier level fusion (intermediate integration), the information is combined within the classifier using separated audio and visual streams, in order to generate a composite classifier to process the individual data streams [9]. This strategy has the advantage of being able to handle possible asynchrony between audio and visual features. In decision level fusion (late integration), independent classifiers are used for each modality and the final decision is computed by the combination of the likelihood scores associated with each classifier [5]. Typically, these scores are fused using a weighting scheme which takes into account the reliability of each unimodal stream. This strategy does not require strictly synchronized streams.

In this paper an isolated word speech recognition system based on audio-visual features is proposed. This system is based on the combination of early and late fusion schemes. In particular, acoustic information is represented by mel-frequency cepstral coefficients, and visual information is represented by coefficients related to mouth shape. The efficiency of the system is evaluated considering noisy conditions in the acoustic channel. The proposed system combines three classifiers based on audio, visual and audio-visual information, respectively, in order to improve the recognition rates through a wide range of signal-to-noise ratios (SNRs), taking advantage of each classifier's efficiency at different SNRs ranges. Two audio-visual databases are employed to test the proposed system. The experimental results show that a significant improvement is achieved when the visual information is considered.

The rest of this paper is organized as follows. The description of the proposed system is given in Sect. 2, and the databases used for the experiments are described in Sect. 3. In Sect. 4 experimental results are presented, where the performance of the proposed strategy is analyzed. Finally, some concluding remarks and perspectives for future work are included in Sect. 5.

## 2 Audio-Visual Speech Recognition System

The proposed system aims to improve speech recognition when the acoustic channel is corrupted by noise, which is the usual situation in most applications, by fusing audio and visual features. In this scenario, the efficiency of a classifier based on audio-only information deteriorates as the SNR decreases, while the efficiency of a visual-only information classifier remains constant, since it does not depend on the SNR in the acoustic channel. However, the use of only visual information is usually not enough to obtain relatively good recognition rates. It has been shown in several works in the literature [6,8,9], that the use of audio-visual feature vectors (early integration) improves the recognition rate in the presence of noise in comparison to the audio-only case. An example of this typical behavior is illustrated in Fig. 1, where the recognition rates for audio-only

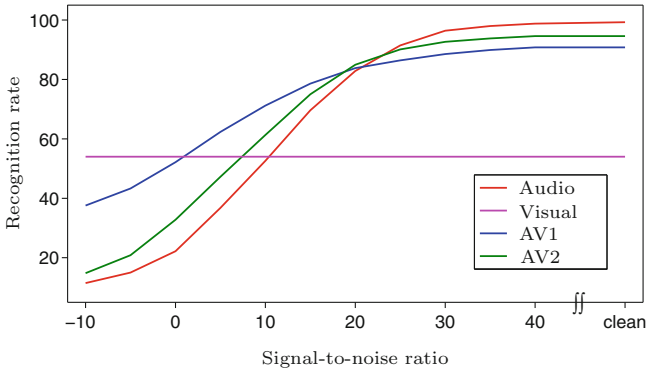



**Fig. 1.** Typical recognition rates for the cases of audio-only, video-only and audio-visual classifiers under acoustic noisy conditions.

(red), visual-only (magenta), and two different audio-visual (blue and green) classifiers as a function of SNR are depicted. These recognition rates were computed using an audio-visual database compiled by the authors. As expected, the audio classifier performs better than the visual one for high SNRs and viceversa. The combination of audio-visual features leads to an improvement of the recognition rates in comparison to the audio-only case. However, for the case of low SNRs, the audio-visual classifier performs worse than the visual one since fused audio-visual features are degraded by the highly corrupted acoustic data. Using different combinations of acoustic and visual features, different performances can be obtained. For instance, if the audio-visual features contain more visual than acoustic information, the performance at low SNRs is improved since visual information is more reliable in this case. However, the efficiency at high SNRs is deteriorated, where the acoustic information is more important. Even for cases where a small portion of audio information is considered, a notorious improvement could be obtained for low SNRs, but the efficiency at high SNRs could be worse than for the audio-only case. Thus, there exists a trade-off between performance at low and high SNRs. These situations are depicted in Fig. 1, where $AV1$ contains more visual information than $AV2$.

Taking into account the previous analysis, the recognition system proposed in this paper combines three different classifiers based on audio, visual and audio-visual information, respectively, aiming at recognizing the input word and maximizing the efficiency over the different SNRs. In the training stage, a combined classifier is trained for each particular word in the vocabulary. Then, given an audio-visual observation sequence associated with the input word to be recognized, denoted as $O_{av}$, which can be partitioned into acoustic and visual parts, denoted as $O_a$ and $O_v$, respectively, the probability $(P_i)$ of the proposed combined classifier corresponding to the $i$-class is given by

$$P_i = P\left(O_a|\lambda_i^a\right)^{\alpha} P\left(O_v|\lambda_i^v\right)^{\beta} P\left(O_{av}|\lambda_i^{av}\right)^{\gamma}, \tag{1}$$

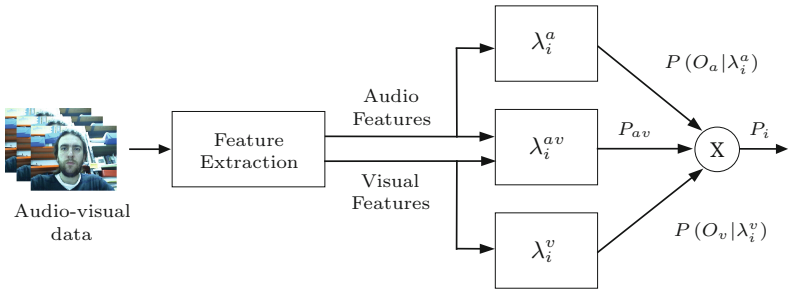



**Fig. 2.** Schematic representation of the computation of the probability associated with a particular class $i$ for the proposed combined classifier. $P_{av}$ refers to $P(O_{av}|\lambda_i^{av})$.

where $P(O_a|\lambda_i^a)$, $P(O_v|\lambda_i^v)$ and $P(O_{av}|\lambda_i^{av})$ are the probabilities corresponding to the audio ($\lambda_i^a$), visual ($\lambda_i^v$) and audio-visual ($\lambda_i^{av}$) classifiers, respectively, and $\alpha$, $\beta$ and $\gamma$ are positive real coefficients that satisfy the following condition

$$\alpha + \beta + \gamma = 1. \tag{2}$$

The visual ($\lambda_i^v$) classifier is more useful at low SNRs ($\beta$ is predominant), where the acoustic data is highly corrupted by noise, while at medium levels of SNRs, the audio-visual classifier ($\lambda_i^{av}$) retrieves better decisions ($\gamma$ is predominant). For high SNR conditions, an audio classifier ($\lambda_i^a$) is employed ($\alpha$ is predominant). A block diagram representing this computation is depicted in Fig. 2.

The audio ($\lambda_i^a$), visual ($\lambda_i^v$) and audio-visual ($\lambda_i^{av}$) classifiers are implemented using left-to-right Hidden Markov Models (HMM) with continuous observations. Audio-visual features are extracted from videos where the acoustic and visual streams are synchronized. The audio signal is partitioned in frames with the same rate as the video frame rate. For a given frame $t$, the first eleven non-DC Mel-Cepstral coefficients are computed and used to compose a vector denoted as $\mathbf{a}_t$. In order to take into account the audio-visual co-articulation, information of $t_a$ preceding and $t_a$ subsequent frames is used to form the audio feature vector at frame $t$, $\mathbf{o}_{at} = [\mathbf{a}_{t-t_a}, \ldots, \mathbf{a}_t, \ldots, \mathbf{a}_{t+t_a}]$, and the information of $t_v$ preceding and $t_v$ subsequent frames is used to form the visual feature vector, $\mathbf{o}_{vt} = [\mathbf{v}_{t-t_v}, \ldots, \mathbf{v}_t, \ldots, \mathbf{v}_{t+t_v}]$, where $\mathbf{v}_t$ contains the visual information at frame $t$. Finally, the audio-visual feature vector is composed by the concatenation of the associated acoustic and visual feature vectors, that is $\mathbf{o}_{avt} = [\mathbf{o}_{at}, \mathbf{o}_{vt}]$, considering $t_a^{av}$ and $t_v^{av}$ frames of co-articulation for the audio and visual features, respectively. Hereafter, the audio, visual and audio-visual classifiers will be denoted as $\lambda_{(s,m)}^{a_{t_a}}$, $\lambda_{(s,m)}^{v_{t_v}}$ and $\lambda_{(s,m)}^{a_{t_a}v_{t_v}}$, respectively, where the subscripts $s$ and $m$ denote the number of states and Gaussian mixtures of the HMM, respectively.

## 3 Audio-Visual Databases

The performance of the proposed audio-visual speech classification scheme is evaluated over two isolated word audio-visual databases, *viz.*, Carnegie Mellon

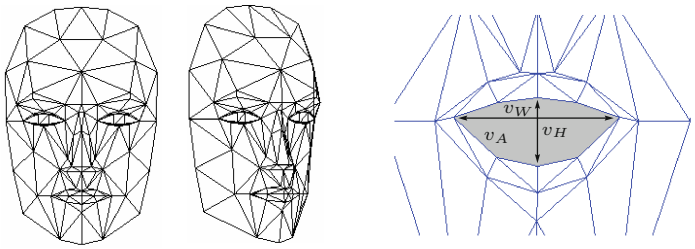


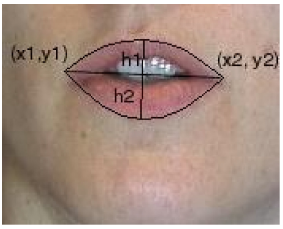

**Fig. 3.** AV-UNR Database visual features. (a) *Candide-3* face model. (b) Visual parameters.

University (AV-CMU) database (now at Cornell University) [2], and a database compiled by the authors, hereafter referred to as AV-UNR database.

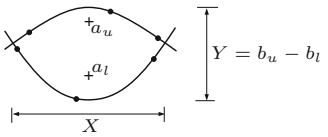
**(I) *AV-UNR* database:** The authors of this paper have compiled an audio-visual database consisting of videos of 16 speakers facing the camera, pronouncing a set of ten words 20 times, in random order. These words correspond to the Spanish utterances of the following actions: *up*, *down*, *right*, *left*, *forward*, *back*, *stop*, *save*, *open* and *close*, a total of 3200 utterances. The videos were recorded at a rate of 60 frames per second with a resolution of $640 \times 480$ pixels, and the audio was recorded at 8 kHz synchronized with the video. Individual words in the database were automatically segmented based on the audio signal, by detecting zero-crossings and energy level in a frame wise basis.

Visual features are represented in terms of a simple 3D face model, namely *Candide-3* [1]. This 3D face model, depicted in Fig. 3(a), has been widely used in computer graphics, computer vision and model-based image-coding applications. The advantage of using the Candide-3 model is that it is a simple generic 3D face model, adaptable to different real faces, that allows to represent facial movements with a small number of parameters. The method proposed by the present authors in [10] is used to extract visual features related to mouth movements during speech. As it is described in [10], this visual information is related to the generic 3D model and it does not depend on the particular face being tracked, *i.e.*, this method retrieves normalized mouth movements. The mouth shape at each frame $t$ is then used to compute three visual parameters, *viz.*, mouth height ($v_H$), mouth width ($v_W$) and area between lips ($v_A$), as depicted in Fig. 3(b). These three parameters are used to represent the visual information at frame $t$.

**(II) *AV-CMU* database:** The AV-CMU database [2] consists of ten speakers, with each of them saying a series of 78 words and repeating the series ten times, resulting in a total of 7800 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 lips ($h2$), as depicted in Fig. 4(a). The visual information was captured with a sample rate of 30 frames per seconds.




(a)




(b)

**Fig. 4.** CMU database. (a) Visual data included in the database. (b) Parabolic lip contour model proposed in [3].

In this paper, the model-based method proposed in [3] is employed to represent the visual information associated to each uttered word. This method is based on weighted least-squares parabolic fitting of the upper and lower lip contours, and it does not require the assumption of symmetry across the horizontal axis of the mouth, and it is therefore more realistic. As described in [3], this model does not depend on the accurate estimation of specific facial points and it is robust to missing or noisy data points. This parameterized lip contour model is based on a pair of intersecting parabolas with opposite orientation, as it is depicted in Fig. 4(b). This parabolic model includes separate parameters for the motion of the upper and lower lips of the mouth during speech. The defining parameters of the model include the focal parameters of the upper and lower parabolas ($a_u$ and $a_l$, respectively) and $X$ and $Y$, the difference between the offset parameters of the parabolas ($b_u$ and $b_l$). As reported in [3], the best representation of the visual information for the AV-CMU database is obtained with a feature vector composed of 5 coefficients, $[Y, X, a_u, a_l, \Theta]$, where

$$
\begin{aligned}
Y &= b_u - b_l \\
X &= 2\sqrt{\frac{b_l - b_u}{a_u - a_l}} \\
\Theta &= \arctan\left\{\frac{\sqrt{(a_l - a_u)(b_u - b_l)}}{2}\right\}
\end{aligned}
$$

Thus, in this paper, these five parameters are used to represent the visual information at each frame of the sequence.

## 4 Experimental Results

The proposed audio-visual speech recognition system is tested separately on the databases described in Sect. 3. To evaluate the recognition rates under noisy acoustic conditions, experiments with additive Gaussian noise, with SNRs ranging from -10 dB to 40 dB, were performed. To obtain statistically significant results, a 5-fold cross-validation (5-fold CV) is performed over the whole data

in each of the databases, to compute the recognition rates. For each instance of the 5-fold CV, audio, visual and audio-visual HMM models are is trained for each word in the database, using the corresponding training set of 5-fold CV. It is important to note that, all the speakers are equally represented in both the training and the testing sets. This evaluation setup corresponds to the so-called "semi-speaker-dependent" approach [11], since both the training and testing sets include the utterances from all speakers.

The three classifiers in the proposed system, based on audio, visual and audio-visual information, respectively, are implemented using left-to-right Hidden Markov Models (HMM) with continuous observations. The tuning parameters of this system are the ones associated with the structure of each HMM classifier, the co-articulation times considered to compose the audio, visual and audio-visual feature vectors, and the coefficients $\alpha$, $\beta$ and $\gamma$ of the decision level integration stage. In order to select the optimum parameters for the classifiers, several experiments were performed considering number of states in the range from 3 to 15, number of Gaussian mixtures from 4 to 20, full covariances matrices, and co-articulation parameters in the range from 0 to 7. Regarding the coefficients $\alpha$, $\beta$ and $\gamma$, which modify the contribution to the final decision of the audio, visual and audio-visual classifiers, respectively (see Eq. (1)), several experiments were performed using different possible combinations of them. In order to obtain better recognition rates over the different SNRs, the values of these coefficients should be modified for the different SNRs, so that the higher contribution at low SNR comes from the visual classifier, at medium SNRs from the audio-visual classifier, and at high SNRs from the audio classifier.

**(I) *AV-UNR* database:** Fig. 5(a) depicts the results, using a boxplot representation, of the evaluation of different configurations for the visual classifier. For each $t_v$, the results associated with the best HMM structure are presented. As it is customary, the top and bottom of each box are the 75th and 25th percentiles of the samples, respectively, and the line inside each box is the sample median. It must be noted that there is no need to carry out this test considering different SNRs, since the visual features are not affected by the acoustic noise. The higher accuracy was obtained for an HMM with 8 states, 17 Gaussian mixtures, and $t_v = 5$, which corresponds to a visual feature vector $\mathbf{o}_{vt}$ composed by 33 parameters, associated to a sliding window of 183 ms in the time domain.

In Fig. 5(b), the results of the experiments to select the proper values for the audio classifier are depicted. These experiments were performed considering several SNRs for the additive Gaussian noise. In this case, only the medians for each noise level are depicted for visual clarity reasons. Although this figure shows the results for a wide range of SNRs, it must be noted that the selection of $t_a$ should be done taking into account that the contribution of the audio classifier to the final decision stage is more important at high SNR conditions. For that reason, an HMM with 3 states and 4 Gaussian mixtures, using $t_a = 5$ is the best option for this classifier.

For the case of the audio-visual classifier, two co-articulation parameters are involved $t_a^{av}$ and $t_v^{av}$. Figure 5(c) shows the recognition rates obtained for three

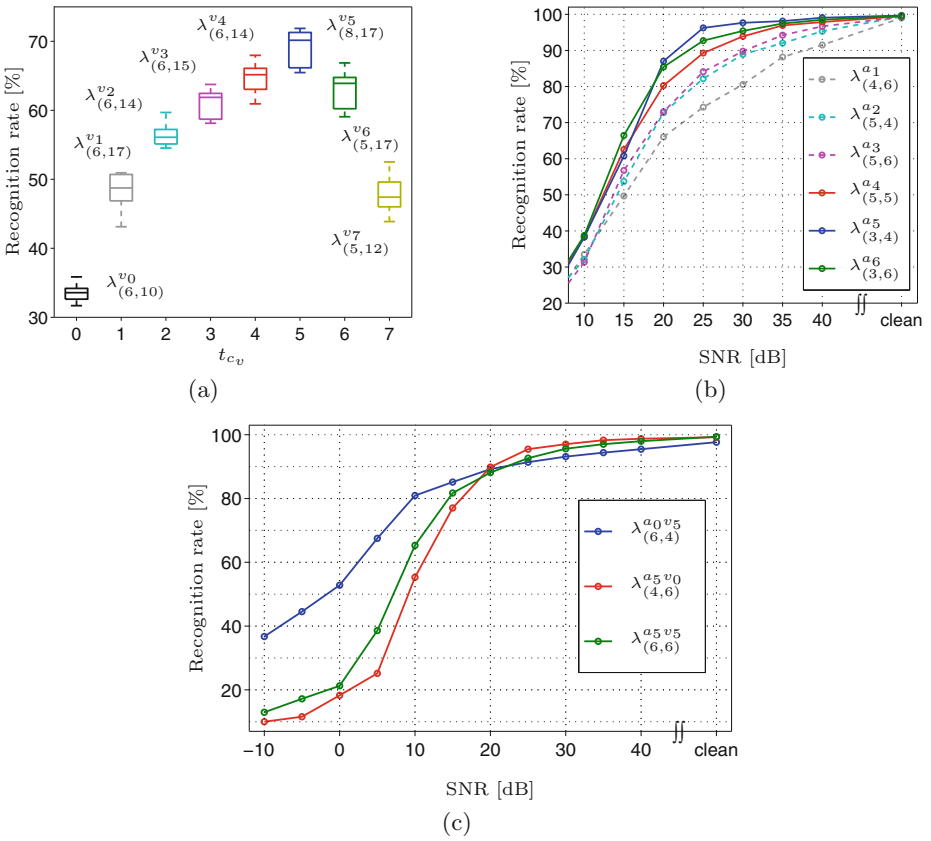



**Fig. 5.** Recognition rates for the (a) visual, (b) audio and (c) audio-visual classifiers using different tuning parameters.

particular audio-visual features configurations, namely $\lambda^{a_0v_5}_{(6,4)}$, $\lambda^{a_5v_5}_{(6,6)}$ and $\lambda^{a_5v_0}_{(4,6)}$, where the number of states and Gaussian mixtures have been optimized for each case. It can be noted from Fig. 5(c) that the best performance at middle SNRs is obtained for the case of configuration ($t^{av}_a = 0, t^{av}_v = 5$), while configurations ($t^{av}_a = 5, t^{av}_v = 5$) and ($t^{av}_a = 5, t^{av}_v = 0$) present a better performances at high SNRs. The performance of the remaining possible configurations lies between upper and lower limiting curves, following the same properties. These results support the comments in Sect. 2, regarding the fact that configurations that use more visual information perform better at low SNRs and viceversa. Regarding the selection of the optimal audio-visual classifier configuration to be used at the final decision stage, it must be taken into account that the contribution of this classifier is important at low and middle range SNR conditions, since at high SNR the audio classifier provides more accurate decisions. Thus, an adequate configuration for this purpose is ($t^{av}_a = 0$ and $t^{av}_v = 5$).

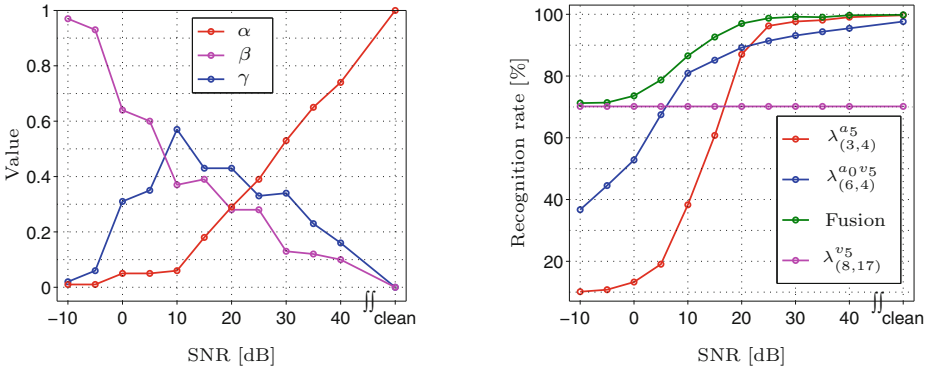



**Fig. 6.** (a) Optimum values for coefficients $\alpha$, $\beta$ and $\gamma$ over the SNRs. (b) Recognition rates of the proposed fusing strategy (green) using the optimum values for the weighting coefficients $\alpha$, $\beta$ and $\gamma$. Performances of audio (red), visual (magenta) and audio-visual (blue) classifiers are also shown (Color figure online).

At this point, the parameters associated with the three classifiers have been selected, and the optimal values for $\alpha$, $\beta$ and $\gamma$ must be chosen. The results of tests performed for this purpose are depicted in Fig. 6(a). As expected, it can be seen that the optimum value of $\alpha$ is the lower one at low SNRs, and it increases as the SNR increases, becoming the higher one at high SNRs. On the other hand, the optimum values of coefficient $\beta$ present an inverse evolution. While for the case of coefficient $\gamma$ the higher values are at medium SNRs.

Figure 6(b) shows the recognition rates obtained with the proposed fusion strategy (green) over the SNRs, using the optimum values for the weighting coefficients $\alpha$, $\beta$ and $\gamma$, presented in Fig. 6(a). In this figure, the recognition rates corresponding to the audio (red), visual (magenta) and audio-visual (blue) classifiers are also depicted. It is clear that the proposed objective of improving the recognition rates through the different SNRs has been accomplished. In addition, the performance of the proposed system is comparable to that of other methods presented in the literature [9]. In these experiments, the SNR of each speech signal was *a priori* known since the noise was intentionally injected in order to evaluate the proposed system at different SNRs. In practical applications, the SNR present in a speech signal can be estimated by comparing its energy with the one corresponding to a previously recorded background noise. A sample of the background noise could be automatically extracted from the silence interval preceding the occurrence of the speech, or it could be recorded on demand by the user. The weights could then be selected from the curves in Fig. 6(a).

**(II) *AV-CMU* database:** The proposed recognition system was also evaluated over the public AV-CMU database [2]. In particular, in order to compare the performance of the proposed system with the one presented in [3], this evaluation

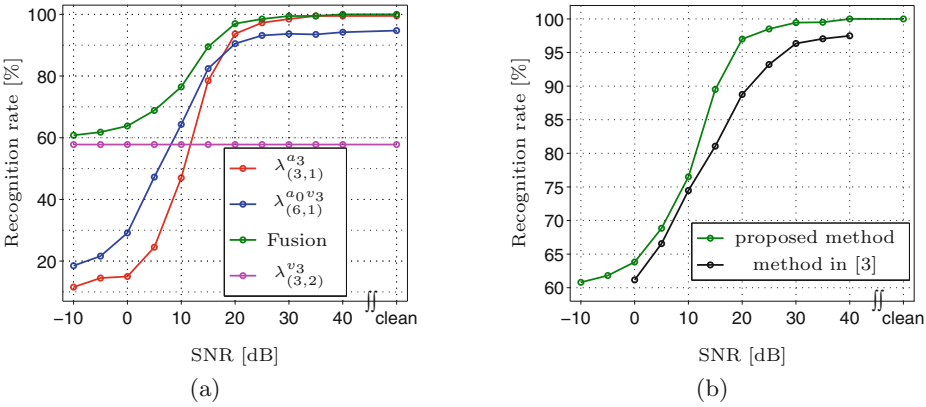



**Fig. 7.** (a) Performance of the proposed recognition system over the AV-CMU database. (b) Efficiency comparison with the method proposed in [3].

was carried out over a subset of ten words, the numbers from 1 to 10. To select the values of the tuning parameters of the system, the same procedure used with the AV-UNR database was employed. The details have not been included due to space limitations. In Fig. 7(a), the performance of the proposed fusion strategy (green) is depicted, where it can be noted that it enforces a significant improvement of the recognition rates through a wide range of SNRs. Figure 7(b) compares the performances obtained with the proposed method and with the one described in [3], evaluated over the same database. It is clear that the proposed method outperforms the one in [3] across all the considered SNRs.

## 5 Conclusions

An isolated word speech recognition system based on audio-visual information was proposed in this paper. This system is based on the combination of early and late fusion schemes. Three classifiers based on audio, visual and audio-visual information, respectively, are combined in order to improve the recognition rates through a wide range of signal-to-noise ratios. The performance of the proposed recognition system was evaluated over two isolated word audio-visual databases. Experimental results show that the structure of the proposed system leads to a significant improvement of the recognition rates through a wide range of signal-to-noise ratios. It is important to note that, the absolute recognition rates could be further improved by considering well-known strategies usually employed in speech recognition, for instance, by incorporating delta mel-cepstral coefficients to the audio features, by including noisy features in the training stage, etc.

## References


1. Ahlberg, J.: Candide-3 - an updated parameterised face. Department of Electrical Engineering, Linköping University, Sweden, Technical report (2001)
2. AMP Lab.: Advanced Multimedia Processing Laboratory. Cornell University, Ithaca, NY, http://chenlab.ece.cornell.edu/projects/AudioVisualSpeechProcessing (Last visited: October 2014)
3. Borgström, B., Alwan, A.: A low-complexity parabolic lip contour model with speaker normalization for high-level feature extraction in noise-robust audiovisual speech recognition. IEEE Trans. Syst. Man Cybern. Part A Syst. Hum. **38**(6), 1273–1280 (2008)
4. Dupont, S., Luettin, J.: Audio-visual speech modeling for continuous speech recognition. IEEE Trans. Multimedia **2**(3), 141–151 (2000)
5. Estellers, V., Gurban, M., Thiran, J.: On dynamic stream weighting for audio-visual speech recognition. IEEE Trans. Audio Speech Lang. Process. **20**(4), 1145–1157 (2012)
6. Jaimes, A., Sebe, N.: Multimodal human-computer interaction: a survey. Comput. Vis. Image Underst. **108**(1–2), 116–134 (2007)
7. Papandreou, G., Katsamanis, A., Pitsikalis, V., Maragos, P.: Adaptive multimodal fusion by uncertainty compensation with application to audiovisual speech recognition. IEEE Trans. Audio Speech Lang. Process. **17**(3), 423–435 (2009)
8. Potamianos, G., Neti, C., Gravier, G., Garg, A.: Recent advances in the automatic recognition of audio-visual speech. Proc. IEEE **91**(9), 1306–1326 (2003)
9. Shivappa, S., Trivedi, M., Rao, B.: Audiovisual information fusion in human computer interfaces and intelligent environments: a survey. Proc. IEEE **98**(10), 1692–1715 (2010)
10. Terissi, L., Gómez, J.: 3D head pose and facial expression tracking using a single camera. J. Univ. Comput. Sci. **16**(6), 903–920 (2010)
11. Zhao, G., Barnard, M., Pietikäinen, M.: Lipreading with local spatiotemporal descriptors. IEEE Trans. Multimedia **11**(7), 1254–1265 (2009)