Audio-Visual Speech Recognition Scheme Based on Wavelets and Random Forests Classification

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

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

Abstract. This paper describes an audio-visual speech recognition system based on wavelets and Random Forests. Wavelet multiresolution analysis is used to represent in a compact form the sequence of both acoustic and visual input parameters. Then, recognition is performed using Random Forests classification using the wavelet-based features as inputs. The efficiency of the proposed speech recognition scheme is evaluated over two audio-visual databases, considering acoustic noisy conditions. Experimental results show that a good performance is achieved with the proposed system, outperforming the efficiency of traditional Hidden Markov Model-based approaches. The proposed system has only one tuning parameter, however, experimental results also show that this parameter can be selected within a small range without significantly changing the recognition results.

Keywords: Speech recognition · Audio-visual speech · Random forests · Wavelet analysis

1 Introduction

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 [9]. 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. This correlation between the acoustic and visual information during speech has motivated, in the last decades, several research activities associated with audio-visual speech recognition [14]. This research has demonstrated that recognition rates in noisy acoustic conditions can be significantly improved in comparison with only-acoustic recognition systems [12].

For audio-visual speech recognition, several kinds of pattern recognition methods have been adopted in the literature such as Linear Discriminant Analysis, Artificial Neural Networks (ANN) [12], matching methods utilizing dynamic programming, K-Nearest Neighbors (K-NN) algorithms [13], Support Vector

A. Pardo and J. Kittler (Eds.): CIARP 2015, LNCS 9423, pp. 567–574, 2015.
DOI: 10.1007/978-3-319-25751-8_68

Machine classifiers (SVM) [16] and Hidden Markov Models [7][8][11]. The most widely used classifiers are traditional HMMs that statistically model transitions between the speech classes and assume a class-dependent generative model for the observed features. In general, these recognition systems require a calibration stage to tune the parameters of the classifier in order to obtain an adequate performance in the recognition. 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 visual features data set being employed.

In this paper, a novel audio-visual speech classification scheme based on wavelets and Random Forests (RF) [4] is proposed. Wavelet multiresolution analysis is used to model the sequence of audio and visual parameters. The coefficients associated with these representations are used as features to model the audio-visual speech information. Speech recognition is then performed using these wavelet-based features and a Random Forests classification method. Random Forests [4] have very good discriminative capabilities, run efficiently on large databases, can handle thousands of input variables avoiding the need for variable selection, are fast and can grow as many trees as it is necessary without overfitting. These good characteristics are inherited by the proposed audio-visual recognition scheme. The performance of the proposed speech classification scheme is evaluated over two different isolated word audio-visual databases.

The rest of this paper is organized as follows. In section 2 the proposed classification scheme for audio-visual speech recognition is presented. The different visual databases employed to evaluate the proposed system are described in section 3. In section 4 experimental results are presented, and the accuracy of the proposed method is compared to the corresponding to traditional Hidden Markov Model-based approaches, over the same databases. Finally, some concluding remarks are given in section 5.

2 Proposed System

A schematic representation of the proposed speech classification scheme is depicted in Fig. 1. In a first stage, Discrete Wavelet Transform (DWT) is applied to the input parameters. The idea is to perform a multilevel decomposition of the time varying input parameters using the DWT and then use the approximation coefficients to represent them. Resampling of the time functions, prior to the DWT decomposition, is needed in order to have a fixed-length feature vector. In this way, independently of the number of frames associated with each word, a resulting fixed length feature vector is obtained. To have a fixed-length feature vector represents an advantage since it makes the comparison between two feature vectors easier. This method is also independent of the kind of input, in this paper the method is evaluated using acoustic and fused audio-visual input parameters. The wavelet-based feature vector computation scheme is depicted in Fig. 2.

In the wavelet decomposition block, a multilevel decomposition of the time functions is performed, and only the approximation coefficients are retained to

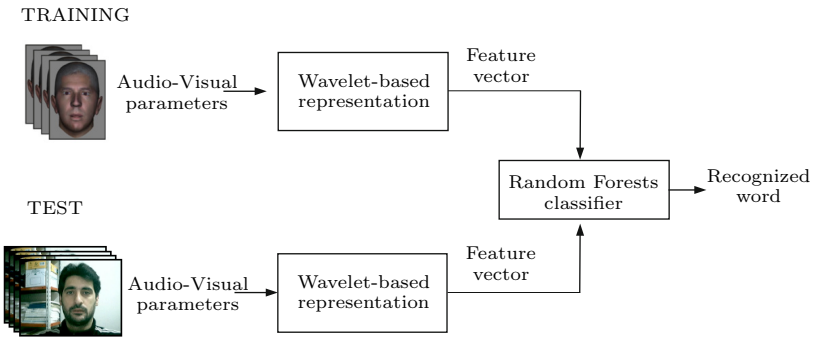


Fig. 1. Schematic representation of the proposed audio-visual speech classification system.

represent them. The approximation accuracy is determined by the chosen resolution level, which also determines the length of the resulting feature vector. Since this length has to be kept reasonably small, there will be a trade-off between accuracy and feature vector length. The design parameter is then the length of the feature vector, which determines the resolution level to be used. The widely used db4 wavelet [5] is employed for the representation of the time functions.

In the second stage, a classification based on Random Forests (RF) is performed. Random Forests is an ensemble of decision trees. The ensemble construction strategy is focused in increasing the diversity among the trees. Decision trees are very unstable (generally a small change in the dataset results in large changes in the developed model [3]), then the diversity among the trees in the ensemble 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. For each node the method selects a small random subset of P attributes (from the total number of attributes available) and use only this subset to search for the best split. The combination of these two sources of diversity produces an ensemble with good prediction performance. This 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 parameters to adjust for 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, 1000 trees are enough). Then, in practice, the only tuning parameter of the model is the number of randomly selected splitting variables to be considered at each node.

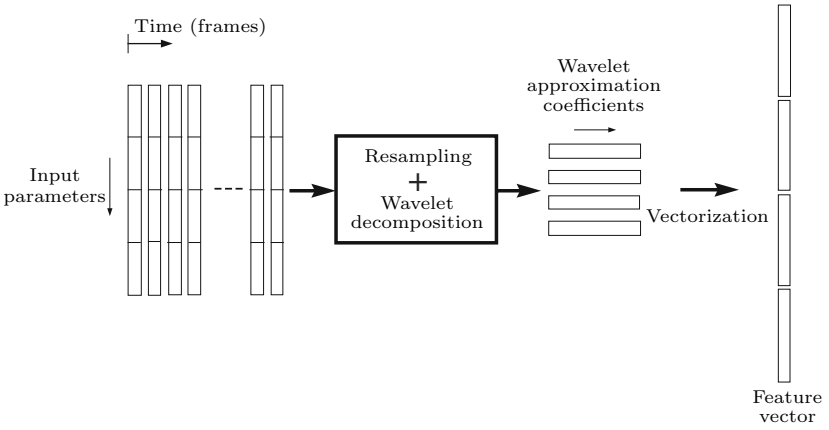


Fig. 2. Schematic representation of the method proposed for computing the wavelet-based feature vector. In this example, the input features are composed by 4 parameters.

3 Audio-Visual Databases

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

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

II) *AV-CMU* database: The AV-CMU database [1] consists of ten speakers, with each of them saying the digits from 0 to 9 ten times. The audio features are represented by the same parameters as in AV-UNR database. To represent the visual information, the weighted least-squares parabolic fitting method proposed in [2] is employed in this paper. Visual features are represented by five parameters, *viz*, the focal parameters of the upper and lower parabolas, mouth's width and height, and the main angle of the bounding rectangle of the mouth.

4 Experimental Results

The proposed audio-visual speech recognition system is tested separately on the databases described in section 3. To evaluate the recognition rates under noisy acoustic conditions, experiments with additive Babble noise, with SNRs ranging from -10 dB to 40 dB, were performed. Multispeaker or Babble noise environment

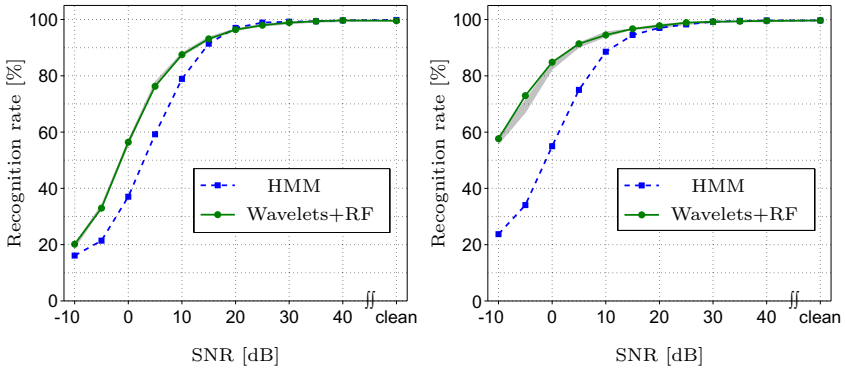
is one of the most challenging noise conditions, since the interference is speech from other speakers. This noise is uniquely challenging because of its highly time evolving structure and its similarity to the desired target speech [10]. In this paper, Babble noise samples were extracted from *NOISEX-92* database, compiled by the Digital Signal Processing (DSP) group at Rice University [6]. To obtain statistically significant results, a two nested 5-fold cross-validation (CV) is performed over the whole data in each of the databases, to compute the recognition rates.

For each database, the evaluation is carried out considering speech data represented by only-audio information on one side and by fused audio-visual information on the other, resulting in four different experiments. Independently of the database being considered, 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 the case of considering audio-visual information, the audio-visual feature vector at frame t is composed by the concatenation of the acoustic parameters with the visual ones.

The tuning parameters of the system are the ones associated with the audio-visual 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 the experiments over the two databases presented in this paper, these parameters are 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. 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 1000 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, hereafter denoted as α.

4.1 Results

The recognition rates of the proposed method over the two audio-visual databases are presented in this subsection. In addition, these results are compared with the ones obtained with a speech recognition system based on Hidden Markov Models (HMMs) over the same databases. Hidden Markov Model approaches have been extensively proposed in the literature for speech recognition [8], and proved to be highly efficient for this task, even on noisy conditions [15]. For comparison purposes, the performance of the HHMs based recognition system was computed for each database, using also two nested 5-fold cross-validation. In particular, the HMMs were implemented using N-state left-to-right models and considering continuous symbol observation, represented by the linear combination of M Gaussian distributions.



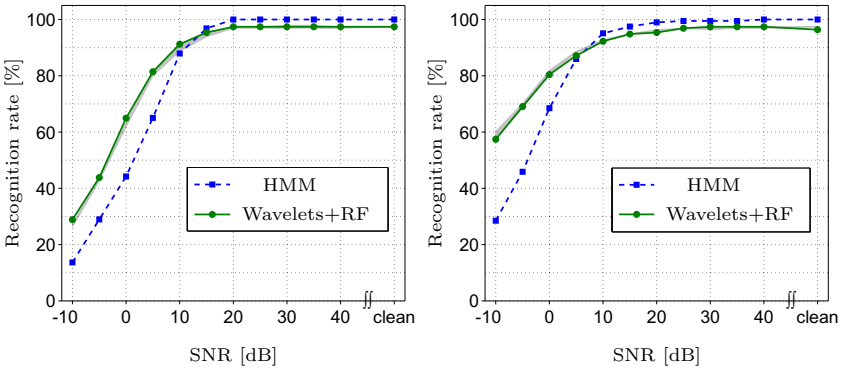
(a) Audio information (b) Audio-visual information

Fig. 3. Recognition rates for different SNRs for the cases of considering (a) only-audio and (b) fused audio-visual information over the AV-UNR database. The performance of the proposed recognition scheme is depicted in solid line (green), while the corresponding to the HMM-based approach is in dashed line (blue). The grey area corresponds to the performances obtained by selecting α in the range from 2 to 6.

AV-UNR database: The results obtained for the cases of considering speech data represented by only-acoustic and audio-visual information over the AV-UNR database are depicted in Fig. 3(a) and 3(b), repectively. It can be seen that for this database, with the proposed approach (solid green line) satisfactory results are obtained, outperforming the HMM-based classification approach (dashed blue line).

AV-CMU database: In Fig. 4(a) and 4(b), the recognition rates obtained over the AV-CMU database are depicted for the cases of considering audio-only and audio-visual information, respectively. As expected, the performance in the recognition task deteriorates as the SNR decreases. For both cases, it can be seen that with the proposed classification scheme (solid green line) good performance is achieved. In particular, in comparison with traditional HMMs approach (dashed blue line), the proposed method leads to significant improvements in the recognition rates for middle and low range SNRs.

The results depicted in Fig. 3 and 4 shows that the proposed method performs well, yielding better performance in comparison with HMM-based methods. As stated before, these experiments were performed by selecting the value for the only tuning parameter, that is the number of randomly selected splitting variables to be considered at each node, via validation procedure (inner CV). However, these experiments also show that this parameter can be selected within a range without significantly affecting the performance of the recognition task. This situation is depicted in Fig. 3 and 4, where the gray areas corresponds to the performances of the system when using α in the range from 2 to 6. Thus, these results indicate that the system can be employed using a fixed setup, *i.e.*, the same wavelet-based representation and RF classifier parameters in all the



(a) Audio information (b) Audio-visual information

Fig. 4. Recognition rates for different SNRs for the cases of considering (a) only-audio and (b) fused audio-visual information over the AV-CMU database. The performance of the proposed recognition scheme is depicted in solid line (green), while the corresponding to the HMM-based approach is in dashed line (blue). The grey area corresponds to the performances obtained by selecting α in the range from 2 to 6.

experiments, and the performance will be similar to the one obtained through a tuning stage. This is an important advantage of the proposed approach in comparison to other methods that necessarily require a usually time consuming optimization stage of the classifiers' metaparameters.

5 Conclusion

An audio-visual speech classification scheme based on wavelets and Random Forests have been proposed in this paper. The sequences of input acoustic and visual parameters are represented via wavelet multilevel decomposition, where only the approximation coefficients are retained to represent them. The proposed representation method leads to a fixed length feature vector, independently of the number of frames associated with each word. This method is also independent of the kind of input feature, either audio-only or fused audio-visual, being considered. These fixed-length wavelet-based feature vectors are then used to model the speech information. Speech recognition is then performed using these wavelet-based features and a Random Forests classification method. The performance of the proposed recognition scheme is evaluated over two different isolated word audio-visual databases. Experimental results show that a good performance is achieved with the proposed system, outperforming the efficiency of traditional Hidden Markov Model-based approaches. The proposed system has only one tuning parameter which can be selected within a small range without significantly changing the recognition results. The experimental results show that the system can be employed using a fixed setup, *i.e.*, the same wavelet-based representation and RF classifier parameters in all the experiments, and the performance will

be similar to the one obtained through a tuning stage. This is an important advantage of the proposed approach in comparison to other methods that necessarily require a usually time consuming optimization stage of the classifiers' metaparameters.

References

1. Advanced Multimedia Processing Laboratory. Cornell University, Ithaca, NY. http://chenlab.ece.cornell.edu/projects/AudioVisualSpeechProcessing/
2. Borgström, B., Alwan, A.: A low-complexity parabolic lip contour model with speaker normalization for high-level feature extraction in noise-robust audiovisual speech recognition. IEEE Transactions on Systems, Man and Cybernetics **38**(6), 1273–1280 (2008)
3. Breiman, L.: Bagging predictors. Machine Learning **26**(2), 123–140 (1996)
4. Breiman, L.: Random forests. Machine Learning **45**(1), 5–32 (2001)
5. Daubechies, I.: Ten Lectures on Wavelets. SIAM, Pennsylvania (1992)
6. NOISEX-92 database. Digital Signal Processing (DSP) group. Rice University, Houston, TX
7. Dupont, S., Luettin, J.: Audio-visual speech modeling for continuous speech recognition. IEEE Trans. Multimedia **2**(3), 141–151 (2000)
8. Estellers, V., Gurban, M., Thiran, J.: On dynamic stream weighting for audio-visual speech recognition. IEEE Transactions on Audio, Speech, and Language Processing **20**(4), 1145–1157 (2012)
9. Jaimes, A., Sebe, N.: Multimodal human-computer interaction: A survey. Computer Vision and Image Understanding **108**(1–2), 116–134 (2007)
10. Krishnamurthy, N., Hansen, J.: Babble noise: Modeling, analysis, and applications. IEEE Transactions on Audio, Speech, and Language Processing **17**(7), 1394–1407 (2009)
11. Papandreou, G., Katsamanis, A., Pitsikalis, V., Maragos, P.: Adaptive multimodal fusion by uncertainty compensation with application to audiovisual speech recognition. Transactions on Audio, Speech, and Language Processing **17**(3), 423–435 (2009)
12. Potamianos, G., Neti, C., Gravier, G., Garg, A., Senior, A.W.: Recent advances in the automatic recognition of audio-visual speech. Proceedings of the IEEE **91**, 1306–1326 (2003)
13. Shin, J., Lee, J., Kim, D.: Real-time lip reading system for isolated korean word recognition. Pattern Recognition **44**(3), 559–571 (2011)
14. Shivappa, S., Trivedi, M., Rao, B.: Audiovisual information fusion in human computer interfaces and intelligent environments: A survey. Proceedings of the IEEE **98**(10), 1692–1715 (2010)
15. Terissi, L.D., Sad, G., Gómez, J.C., Parodi, M.: Noisy speech recognition based on combined audio-visual classifiers. Lecture Notes in Computer Science **8869**, 43–53 (2015)
16. Zhao, G., Barnard, M., Pietikäinen, M.: Lipreading with local spatiotemporal descriptors. IEEE Transactions on Multimedia **11**(7), 1254–1265 (2009)