Audio-Visual Speech Classification based on Absent Class Detection

Gonzalo Daniel Sad Juan Carlos Gómez
CIFASIS, CONICET, and FCEIA, Universidad Nacional de Rosario, Argentina
Email: {sad, gomez}@cifasis-conicet.gov.ar

***Abstract*—In the present paper, a novel method for Audio-Visual Speech Recognition is introduced, aiming to minimize the intra-class errors. Based on a novel training procedure, the Complementary Models are introduced. These models aim to detect the absence of a class, in contrast to traditional models that aim to detect the presence of a class. In the proposed method, traditional models are employed in the first stage of a cascade scheme, and then the proposed complementary models are used to make the final decision on the recognition results. Experimental results in all the scenarios evaluated (different inputs modalities, three databases, four classifiers, and acoustic noisy conditions), show that a good performance is achieved with the proposed scheme. Also, better results than other reported methods in the literature over two public databases are achieved.**

***Index Terms*—Audio-Visual Speech Recognition, Complementary Models, Ensemble Models.**

I. Introduction

It is of common knowledge that, besides the acoustic signal, the visual information during speech related to facial expressions, hand gesture and body posture contributes significantly to the intelligibility of the message being transmitted, and to the perception of the actual meaning of the message [1]. In addition, as pointed out in a recent survey about the interaction between gesture and speech [2], the parallel use of these modalities gives the listener access to complementary information not present in the acoustic signal by itself. In recent years, the study of human communication has benefited from the increasing number of multimodal corpora available to researchers in this field. Significant research effort has been devoted to the development of Audio Visual Speech Recognition Systems (AVSRS) where the acoustic and visual information (mouth movements, facial gestures, etc.) during speech are taken into account [3].

Several methods and models have been developed to perform speech recognition by fusioning audio and visual information [3]. One of the first methods presented in the literature (and one still widely used) is Hidden Markov Models (HMMs). Others classical methods from the Machine Learning area has been also implemented, like Adaptive Boosting classifiers (AdaBoost), Support Vector Machine (SVM), Random Forests (RF), etc. In recent years, more sophisticated methods like Restricted Boltzmann Machines (RBM), Deep Learning (DL) and sparse coding, have proved to be very suitable for speech recognition tasks.

For several decades, the research on model combination or ensemble of models has been an active area in the Artificial Intelligence and Machine Learning fields [4], [5]. The different approaches to build complementary models available in the literature, can be divided into two major categories: ad-hoc methods and explicit methods. In the first category, the models of the ensemble are generated altering some variable or algorithm of the model, looking for some diversity among them so that each model delivers different errors [6]. Some examples of this category are: altering the optimization algorithm, using different input features [7], bootstrapping the training data, etc. In the second category, the diversity among the models in the ensemble is achieved explicitly in the training procedure [8], [9] by training all the models in parallel or in an iteratively fashion. This latest approach usually yields an overfitted model, due to the excessive complexity injected in the training stage. Different works in the literature of AVSRS have proposed ensemble models where the complementary models are generated mostly based in ad-hoc methods [3].

In the present paper, a novel scheme for speech classification tasks based on the combination of traditional and complementary models, is proposed to improve recognition rates. The proposed scheme aims to minimize the intra-class errors. The diversity among the models is achieved by means of an ad-hoc method, altering the way the training data is employed in the training procedure. The proposed scheme can be implemented with different models, *viz.*, generative models like HMM or RBM and discriminative models like RF, SVM, AdaBoost, etc. There are no restrictions about the kind of input features on the proposed method, *i.e.*, it can be employed for lip-reading tasks, where the inputs are visual features, for audio speech recognition, where the inputs are audio features, and also for audio-visual speech recognition, where the inputs are acoustic and visual features previously fused. Given a model or classifier, the corresponding complementary model proposed in this paper is generated using the same training procedure as in the original model or classifier, but defining a new set of classes for the training examples, aiming to detect absence of a class. In the complementary models, the i-th class is formed using all the instances in the vocabulary except the corresponding to class i. For instance, let consider a vocabulary composed by three classes, C_1, C_2 and C_3, the new classes are defined as AC_1 which contains all the examples of the classes C_2 and C_3, AC_2 which contains all the examples of the classes C_1 and C_3, and AC_3 which contains all the examples of the classes C_1 and C_2.

The remainder of this paper is organized as follows. Sec-

tion II presents a preliminary analysis of the class confusability problem and Section III introduces and explains the classification based on absent classes. In section IV, a description of the proposed system is given. The databases used for the experiments are described in Section V, and the experimental results and the performance of the proposed strategy is analyzed in Section VI. Finally, some concluding remarks are given in Section VII.

II. Intra-class classification problem detection

Phonemes and words are usually the fundamental units to be recognized in any speech recognition system, and they are represented by different classes in the implemented classifier. As in any classification task, the classifier could make mistakes, which in principle could be considered as random. But in speech recognition tasks it is very usual to find out that these classification errors are mostly due to a certain set of classes. This is known in the literature as intra-class confusion or Class confusability. In lip-reading tasks (visual speech recognition), most errors are due to confusion between speech utterances consisting of identical sequence of visemes (a viseme is the visual equivalent of a phoneme). As an example, let's suppose the case of being detecting the letters of the alphabet [A-Z] by lipreading. The letters B and P will be the responsible of most classification errors because even though they are defined by different phonemes ([B+IY] and [P+IY], respectively), they are mapped to the same visemes: /p+iy/. A similar situation arises for the letters T, C and D. Even though they are defined by different phonemes ([T+IY], [C+IY], and [D+IY], respectively), they are mapped to the same visemes: /t+iy/. They will also be a significant source of classification errors.

Such kind of behavior could be detected by a simple inspection of the confusion matrix obtained in the classifier's evaluation stage. The confusion matrix has p rows (reflecting the classifier output decision) and p columns (reflecting the true classes), where p represents the number of classes. Ideally, all the off-diagonal elements of the confusion matrix will be 0, which indicates that no classification errors were made. But, as stated before, in speech recognition tasks usually there exist some errors due to a certain set of classes, which is reflected as high-value off-diagonal elements in the confusion matrix. The rest off-diagonal elements usually remain with low values. Therefore, the most confused classes can be determined by the position (column and row) of these off-diagonal elements with high-values. In the method proposed in this paper, the P off-diagonal elements with the highest values will determine (by its column and row positions) the selected conflicted classes.

III. Complementary models based on the absent class classification

In the Machine Learning area, the models and classifiers can be roughly divided into two broad groups, *viz.*, discriminative models (AdaBoost, RF, SVM, etc.) and generative models (naive Bayes, RBM, HMM, etc.). And as it is known, some of these models are used in AVSRS. For the case of generative models, each word, phoneme, viseme, etc., composing the dictionary of classes, is represented by an individual model. In the training stage, these models are trained with examples corresponding to each particular class being representing. Then, each model is evaluated with the given input observation to be recognized (O). Finally, the recognized class is determined by the model giving the highest output probability value. For the case of discriminative models, all the classes composing the dictionary are modeled within a single classifier, which is trained with all the examples. Then, this model is evaluated with the given input observation to be recognized (O) and the recognized class is determined by the one giving the highest probability value. In the present paper, another way of using the data in the training stage is proposed, aiming to detect the absence of a class. This is accomplished by redefining the classes employed by both kinds of models. This approach is different from the traditional training of models described before, since these models aims to detect the presence of a class. The models obtained with this new technique will be called *Complementary Models*. Each original class i in the dictionary is replaced by a new class AC_i, which is defined employing all the examples of all original classes except the corresponding ones to class i. That is,

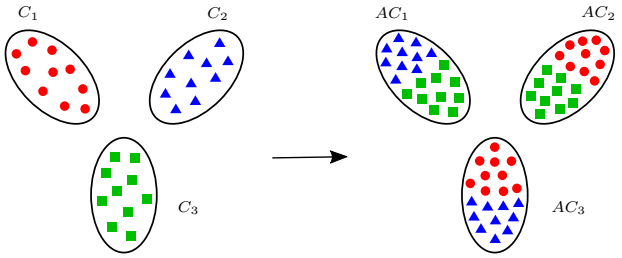


Fig. 1. Procedure to form the new complementary models' classes.

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

where AC_i are the new proposed classes, C_i represents the original classes, I_N represents the N original classes, and $I_N - \{C_i\}$ represents all the N original classes except the C_i class. An example of the procedure to form the new classes for a vocabulary of size $N = 3$, is depicted in Fig. 1.

Then, the models are trained using this new set of classes AC_i. Given an input observation sequence of class i, the recognized class by these new Complementary Models will be determined by the class with the lowest output probability value (AC_i in this case). This seems reasonable, given that these new models detects the absence of a class (ith class in this case). For the case of discriminative model λ, the decision rule of the Complementary Model is

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

where λ_{AC} represents the Complementary Model of λ (trained with the AC_j classes), $P\left(O|\lambda_{AC}, AC_j\right)$ represents the output

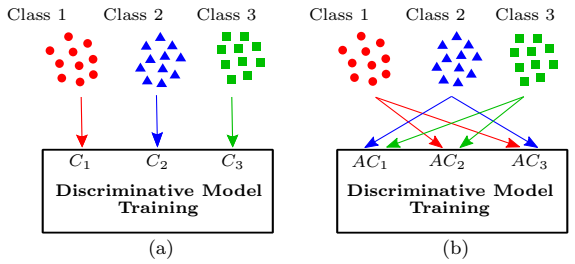


Fig. 2. Training procedure of the proposed complementary models for the discriminative models. (a) Traditional model. (b) Complementary model.

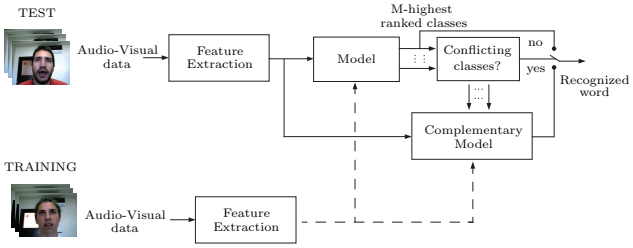


Fig. 3. Schematic representation of the proposed cascade of classifiers.

probability value of the new class AC_j given the observation sequence O and the Complementary Model λ_{AC}, and i represents the recognized class. An example of the training procedure to form both, the discriminative model λ and its corresponding Complementary Model λ_{AC}, for a vocabulary of size $N = 3$, is depicted in Fig. 2.

IV. Proposed Combination of Models

In order to improve recognition rates in speech classification tasks, a novel scheme based on a cascade of classifiers is presented in this section. The proposed scheme, based on the combination of complementary and traditional models, could be implemented with any kind of models (generative or discriminative) and can handle different kinds of input information, *viz.*, visual, audio-visual and audio information. Based on the method proposed in Section III, the complementary model of a given traditional model is obtained and employed in the second stage of the proposed cascade scheme. The complementary model is evaluated only when the proposed system determines that there may be errors in the results of the traditional model due to intra-class confusion. Fig. 3 schematically depict the proposed speech classification cascade scheme.

In the training stage, the traditional model (λ) is trained as usual, employing the audio-visual features extracted from the training data. Also, based on the proposed method in Section II, this trained model is employed to determine the conflicting classes, which are used to train the complementary models. Then, the test stage is carried out in two steps. In the first step, the traditional model is employed to evaluate the input observation. Based on the classification results, the M classes with the highest output probability values are preselected. Then, the previously determined conflicting classes are

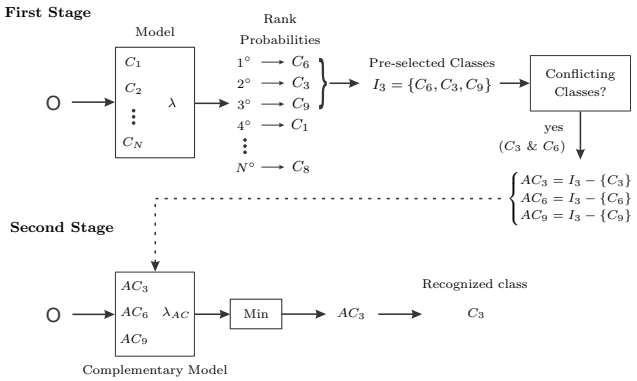


Fig. 4. Example of the proposed classifier combination strategy.

used to detect if there exists a possible case of class confusion between these M classes. This is carried out checking if any pair of these M classes matches any of the conflicting classes. If not, the class with the highest probability determine the recognized word. Otherwise, the method moves to a second step where the complementary model (λ_{AC}) associated with these M classes is evaluated to determine the recognized class. A schematic representation of the proposed cascade of classifiers in the present paper is shown in Fig. 4, considering a discriminative model λ and $M = 3$. First, the traditional model (λ) is evaluated using as input the observation sequence O related to the word to be recognized. Next, the output probability values ($P(O|\lambda, C_i), i = 1, 2, \ldots, N$) are sorted. Then, the complementary model λ_{AC} is selected based on the $M = 3$ highest ranked values ($I_3 = \{C_6, C_3, C_9\}$), which has been previously trained in the training stage. This is because an intra-class confusion case has been detected (C_3 and C_6). Particularly, the new set of classes proposed in this paper is defined as: $AC_9 = I_3 - \{C_9\}$, which is composed of the training examples of original classes C_6 and C_3; $AC_6 = I_3 - \{C_6\}$, which is composed of the training examples of original classes C_9 and C_3; and $AC_3 = I_3 - \{C_3\}$, which is composed of the training examples of original classes C_9 and C_6. After evaluating the complementary model, the recognized word is determined by the new class AC_i associated to the minimum output probability value. In this case: AC_3.

V. Audio-Visual Databases

In order to evaluate the proposed scheme performance, experiments on three audio-visual databases are performed.

I) *AVLetters* database: In this database, 10 speakers repeat three times each of the isolated letters A-Z [14]. Therefore, the complete database consists of the recordings of 780 utterances. The method described in [15] is employed to extract and represents the visual information related to speech. Based on this approach, the motion of the mouth region and the time order in pronunciation of the words being uttered are described be means of spatiotemporal local binary patterns. Hence, a feature vector of 1770 coefficients for each frame image composing this database is obtained.

II) *AV-CMU* database: In this database [10], 10 speakers repeat ten times 78 isolated words commonly used (digits, days, months, names, etc.). In the present paper, only a subset of the vocabulary is employed for the experiments, *viz.*, the numbers from 1 to 10. Therefore, the subset of this database employed in this paper consists of the recordings of 1000 utterances. The method described in [11], based on a weighted least-squares parabolic fitting, is employed to represents the visual information. As a result, a feature vector composed of 5 parameters is obtained, *viz.*, mouths width and height, the main angle of the bounding rectangle of the mouth and the focal parameters of the upper and lower parabolas.

III) *AV-UNR* database: In this database (compiled by the authors of this paper [17]), 16 speakers repeat 20 times 10 isolated words related to movement actions (*close*, *open*, *down*, *up*, etc.). Therefore, the complete database consists of the recordings of 3200 utterances. The method described in [12], based on a 3D face model, namely *Candide-3* [13], is employed to represents the visual information. As a result, a feature vector composed of 3 parameters is obtained, *viz.*, the area between lips and the mouths width and height.

VI. Experimental Results

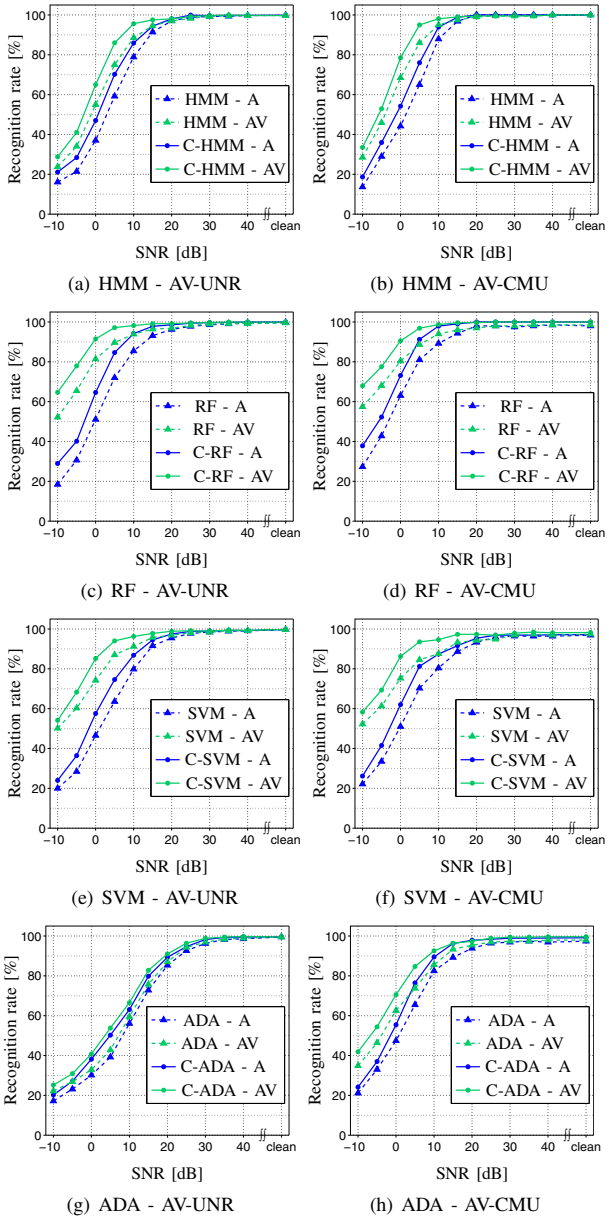
The performance of the proposed scheme is evaluated separately in different scenarios, *viz.*, considering only audio information, only video information (lip-reading), and fused audio-visual information, respectively. These evaluations are carried out over three different databases. In order to evaluate the proposed system in Section IV, four different models are used in both stages of the cascade, *viz.*, Support Vector Machine (SVM), Hidden Markov Model (HMM), Adaptive Boosting (ADA) classifier and Random Forest (RF). 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, adding Babble noise in the audio stream with signal-to-noise ratios (SNRs) ranging from -10 dB to 40 dB. In order to obtain statistically significant results, at each experiment a D-fold crossvalidation (CV) is performed over the whole data to compute the recognition rates. For the cases of AV-CMU and AV-UNR databases, at each fold, one speaker is used for testing and the remaining ones for training, resulting in a speaker independent evaluation (10-fold CV and 16-fold). For the case of the AVLetters database, the evaluation is performed with the same protocol employed in other approaches reported in the literature [15], [16] over this database. The audio features are represented by the first eleven non-DC Mel-Cepstral coefficients, and its associated first and second derivative coefficients. In the audio-visual speech recognition experiments, for each frame, the audio-visual feature vector is obtained concatenating the corresponding audio and visual features. Since SVM, RF and ADA classifiers cannot handle variable length input data, as is the case of speech recognition systems (due to its time varying nature), the proposed method in [17] based on a wavelet feature extraction technique, is employed to obtain fixed-length input data representation. In order to minimize the intra-class errors, the procedure described in Section II is applied to detect the conflicted classes. For each scenario being considered, the classifier (HMM, RF, SVM or ADA) is evaluated randomly splitting each dataset in a train set (70%) and a test set (30%), the confusion matrix is obtained, and the 4 most conflicting pair of classes ($P = 4$) are selected. Then, the cascade of classifiers proposed in Fig. 3 is evaluated, using 3-class complementary models ($M = 3$). The pre-selected P-pairs of conflicted classes are employed in the second stage of the proposed system to determine if the complementary models must be evaluated. Also, several experiments were carried out where the models' (HMM, RF, SVM and ADA) meta-parameters were optimized through an exhaustive search. The recognition rates obtained at different SNRs over the AV-CMU and AV-UNR databases, using audio-only information and fused audio-visual information are depicted in Fig. 5. It is clear that, for both databases, both types of inputs and the four models, the proposed scheme (C-HMM, C-RF, C-SVM and C-ADA) performs better than the ones based on the traditional models (HMM, RF, SVM an ADA). Comparing the results obtained for the audio-only information and fused audiovisual information cases, it can be observed that in general the improvements obtained were greater mainly at low and middle range of SNRs. A maximum improvement of 14% is reached in Fig. 5(c), at SNR = 0dB, for the AV-UNR database using audio-only information and RF. The results for the lip-reading scenario for AVLetters database, are shown in Table I. As it can be observed, the use of the proposed cascade of classifiers scheme (C-HMM, C-RF, C-SVM and C-ADA) improves the recognition rates for all the models been considered. For AVLetters database, a significant improvement of 7% is achieved. Due to limitation space, only some results are showed.

TABLE I
Classification over AVLetters Database considering only visual information.

Classifier	Visual Features	Accuracy
SVM [15]	LSD [15]	58.85 %
Deep Autoencoder [18]	DEEPA [18]	64.40 %
RTMRBM [16]	MRPCA [16]	64.63 %
HMM	LSD [15]	57.30 %
C-HMM	LSD [15]	61.25 %
SVM	LSD [15]	63.08 %
C-SVM	LSD [15]	69.97 %
ADA	LSD [15]	54.23 %
C-ADA	LSD [15]	60.57 %
RF	LSD [15]	65.38%
C-RF	LSD [15]	**72.34%**

VII. Conclusion

In the present paper, a novel scheme composed by a cascade of classifiers, is proposed to improve recognition rates. Traditional models are employed in the first stage of this cascade scheme, and then the proposed complementary models are used to make the final decision on the recognition results.



(a) HMM - AV-UNR (b) HMM - AV-CMU

(c) RF - AV-UNR (d) RF - AV-CMU

(e) SVM - AV-UNR (f) SVM - AV-CMU

(g) ADA - AV-UNR (h) ADA - AV-CMU

Fig. 5. Classification based on audio (A) and fused audio-visual (AV) information. Recognition rates obtained over the AV-UNR (first column) and AV-CMU (second column) databases for different SNRs of Babble noise.

Based on a novel training procedure, the complementary models are introduced. These models aim to detect the absence of a class, in contrast to traditional models that aim to detect the presence of a class. Since the proposed scheme can handle different kinds of input data, the performance evaluation was divided into different scenarios, *viz.*, considering only audio information, only video information, and fused audio-visual information, respectively. To perform this task, three different databases were employed, one compiled by the authors of this chapter and the remaining two are public ones. In order to evaluate the performance of the proposed method against noise injected in the audio stream, experiments with Babble noise were carried out. Based on the experimental results in all the scenarios evaluated, a good performance is achieved with the proposed scheme. Also, better results than other reported methods in the literature over the two public databases are achieved. It is interesting to note that in all the experiments, the proposed method outperforms the traditional ones. Therefore, recognition rate improvements on an existing model could be obtained by resorting to the proposed strategy. To this end, only a training stage of the complementary models is needed, without any modification on the traditional models.

REFERENCES

[1] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, no. 5588, pp. 746–748, 1976.

[2] P. Wagner, Z. Malisz, and S. Kopp, "Gesture and speech in interaction: An overview," *Speech Communication*, vol. 57, pp. 209–232, 2014.

[3] A. K. Katsaggelos, S. Bahaadini, and R. Molina, "Audiovisual fusion: Challenges and new approaches," *Proceedings of the IEEE*, vol. 103, no. 9, pp. 1635–1653, Sept 2015.

[4] T. G. Dietterich, "Ensemble methods in machine learning," in *Proceedings of the First International Workshop on Multiple Classifier Systems*, 2000, pp. 1–15.

[5] H. Liu, Y. Du, and Z. Wu, "Aem: Attentional ensemble model for personalized classifier weight learning," *Pattern Recognition*, vol. 96, p. 106976, 2019.

[6] M. Hwang, W. Wang, X. Lei, J. Zheng, O. Cetin, and G. Peng, "Advances in mandarin broadcast speech recognition," in *Proceedings of the 8th annual conference of the international speech communication association*, 2007, pp. 2613–2616.

[7] M. Koziarski, B. Krawczyk, and M. Woźniak, "The deterministic subspace method for constructing classifier ensembles," *Pattern Analysis and Applications*, vol. 20, no. 4, pp. 981–990, 2017.

[8] O. J. Prieto, C. J. Alonso-González, and J. J. Rodríguez, "Stacking for multivariate time series classification," *Pattern Analysis and Applications*, vol. 18, no. 2, pp. 297–312, 2015.

[9] C. Breslin, "Generation and combination of complementary systems for automatic speech recognition," Ph.D. dissertation, Cambridge University Engineering Department and Darwin College, 2008.

[10] F. J. Huang and T. Chen, "Advanced Multimedia Processing Laboratory. Cornell University, Ithaca, NY," 1998, http://chenlab.ece.cornell.edu/projects/AudioVisualSpeechProcessing, Last visited: February 2020.

[11] B. Borgström and A. Alwan, "A low-complexity parabolic lip contour model with speaker normalization for high-level feature extraction in noise-robust audiovisual speech recognition," *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, vol. 38, no. 6, pp. 1273–1280, 2008.

[12] L. D. Terissi and J. C. Gómez, "3D head pose and facial expression tracking using a single camera," *Journal of Universal Computer Science*, vol. 16, no. 6, pp. 903–920, 2010.

[13] J. Ahlberg, "Candide-3 - an updated parameterised face," Department of Electrical Engineering, Linkping University, Sweden, Tech. Rep., 2001.

[14] I. Matthews, T. Cootes, J. A. Bangham, S. Cox, and R. Harvey, "Extraction of visual features for lipreading," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, p. 2002, 2002.

[15] G. Zhao, M. Barnard, and M. Pietikäinen, "Lipreading with local spatiotemporal descriptors," *IEEE Transactions on Multimedia*, vol. 11, no. 7, pp. 1254–1265, Nov. 2009.

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

[17] L. D. Terissi, G. D. Sad, and J. C. Gómez, "Robust front-end for audio, visual and audio—visual speech classification," *International Journal of Speech Technology*, vol. 21, no. 2, pp. 293–307, 2018.

[18] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Ng, "Multimodal deep learning," in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, 2011, pp. 689–696.