Improving Viseme Recognition using GAN-based Frontal View Mapping

Dário Augusto Borges Oliveira Andrea Britto Mattos
Edmilson da Silva Morais
IBM Research
Rua Tutóia, 1157, Paraíso, São Paulo, Brazil
{dariobo,abritto,edmorais}@br.ibm.com

Abstract

*Deep learning methods have become the standard for Visual Speech Recognition problems due to their high accuracy results reported in the literature. However, while successful works have been reported for words and sentences, recognizing shorter segments of speech, like phones, has proven to be much more challenging due to the lack of temporal and contextual information. Also, head-pose variation remains a known issue for facial analysis with direct impact in this problem. In this context, we propose a novel methodology to tackle the problem of recognizing visemes – the visual equivalent of phonemes – using a GAN to artificially lock the face view into a perfect frontal view, reducing the view angle variability and simplifying the recognition task performed by our classification CNN. The GAN is trained using a large-scale synthetic 2D dataset based on realistic 3D facial models, automatically labelled for different visemes, mapping a slightly random view to a perfect frontal view. We evaluate our method using the GRID corpus, which was processed to extract viseme images and their corresponding synthetic frontal views to be further classified by our CNN model. Our results demonstrate that the additional synthetic frontal view is able to improve accuracy in **5.9%** when compared with classification using the original image only.*

1. Introduction

Visual Speech Recognition (VSR) is the process of interpreting spoken text using video information and is useful when audio data is unavailable or corrupted. Previous studies have demonstrated that audio-video data can improve the recognition of audio-only data in noisy environments [2]. In this context, several works have addressed the recognition of sentences, words, and, to a lesser extent, visemes. A *viseme* is the visual correspondent of a phoneme, *i.e.*, the mouth shape when a user pronounces a phoneme, and given viseme may represent more than one phoneme.

Viseme recognition has valuable characteristics in comparison to the recognition of larger speech segments. For instance, in order to identify words or sentences, a system must be trained using data captured in a specific language or vocabulary; on the contrary, a viseme recognition system may be more easily applied for different languages that share a common set of phonemes – and consequently, visemes. For instance, according to the mapping from [3], the US-English set of visemes encompasses all visemes from the Dutch language and all visemes from Portuguese, Spanish, Italian, and French, except for the viseme representing the phonemes /ɲ/ and /ʎ/ – denoted by the International Phonetic Alphabet (IPA) syntax. Therefore, a training dataset built using visemes instead of words or sentences, can be adapted more easily for different languages without the need to re-train the algorithm with entirely new model and data. Moreover, automatic viseme recognition may be applied to lip-synchronization for synthesizing speech in a video sequence [22].

Recently, synthetic datasets have been applied successfully for the problem of facial analysis by Convolutional Neural Networks (CNNs), more precisely, for facial expression recognition [1] and viseme recognition, using a basic transfer learning approach [20]. However, when it comes to facial image analysis, the variability of the view angle still impacts greatly the performance of classification methods. In this work, we propose to enhance the power of CNN-based automated viseme recognition by using Generative Adversarial Networks (GANs) to artificially generate a synthetic version of a given input mouth shape, perfectly locked in frontal view, and used in addition to the original image for the classification process. To train the GANs we created a large set of 2D synthetic realistic face images obtained from realistic 3D facial models.

This paper is structured as follows. In Section 2, we list previous work on automated VSR focusing on viseme-based approaches. In Section 3, we describe the methodology for mapping input mouth images from an arbitrary

view to a synthetic frontal view and the proposed classification approach. Section 4 describes our experiments with the popular GRID corpus [9] and the obtained results. The paper is concluded in Section 5.

2. Related Work

Several works have addressed different aspects concerning the quality of VSR. Because the correspondence between phonemes and visemes is characterized as one-to-many, some studies have addressed how different phoneme-to-viseme mappings affect automated lip-reading performance; such works are explored in detail by Bear and Harvey [5]. Koumparoulis *et al.* [15] have demonstrated that the design of the mouth ROI (region-of-interest) affects VSR performance significantly: the smallest error rates corresponds to a ROI that includes part of the lower face region, in addition to the mouth, solely.

Formerly, VSR was often addressed by Active Appearance Models (AAMs) and Hidden Markov Models (HMMs), as reviewed extensively by [24]. Recently, Deep Learning-based methods have drastically improved state-of-the-art accuracy for the task. Chung *et al.* [8] use a *'Watch, Listen, Attend and Spell'* (WLAS) network for classifying sentences of the GRID corpus with 97% accuracy. For comparison, the previous baseline for the dataset, using AAMs, was 65% [16].

Another component that should be considered is the size of the speech segment to be recognized. Although sentences from GRID corpus can be identified with very high accuracy, recognizing letters and digits have proven to be a much more challenging task. The baseline for AVLetters dataset [18] on the recognition of isolated letters is 64.63%, using a temporal multi-modal Deep learning-based architecture [12]. The baseline for AVDigits dataset, using the same method, is 40.66% on the recognition of isolated digits.

Regarding even shorter speech segments, in the task of viseme recognition, authors have used SVM to achieve 63% accuracy on distinguishing between 6 viseme classes [21], deep CNNs to obtain 55.7% accuracy on the recognition of 12 viseme classes [14], deep NNs for recognizing 13 viseme classes with 46.61% accuracy [23], and AAMs with a small dataset of two users for achieving around 40-50% accuracy on the recognition of 18 viseme classes [6]. The lower accuracy for short speech segments is expected: previous work demonstrated that human lip-reading performance increases for longer words, indicating the importance of temporal features [4], that also provide valuable data for automated VSR.

In this paper, we describe a novel methodology to recognize visemes with a GAN-based schema for synthetically generating a perfect frontal view of a mouth shape to be used as additional information to a CNN state-of-art classifier.

3. Method

Our methodology is composed by three major steps, as depicted in Figure 1. First, we create a large synthetic database composed by pairs of synthetic images: a random near-frontal image, and the corresponding perfectly frontal image. Then, we train a GAN to map a random near-frontal image into the corresponding perfectly frontal view image. Finally, we use this additional view to train a CNN for classifying visemes. We describe each step in details in the following.

3.1. Viseme Map

Prior works have explored several phoneme-to-viseme mappings in order to assess the most indicated for visual-only computer lip-reading. The map proposed by Lee and Yook [17] is commonly used and has proven to work effectively for this scenario, therefore, it was our choice for this study. The viseme images extracted from the synthetic and real datasets were selected and grouped into 11 classes, with 5 vowels and 6 consonants.

Table 2 displays the grouped phonemes – described in the IPA syntax – and corresponding viseme classes. We denote each viseme class by an identifier to which we will refer in the remaining of the text. In Table 1, each of the 11 classes is represented by a real subject from the GRID dataset and a synthetic subject on frontal and profile views.

3.2. Synthetic Dataset Generation

The GAN used in this work was trained using pairs of mouth images captured at random angles and their corresponding frontal view images. The generation of this dataset is divided in two steps: (i) modeling the faces; and (ii) rendering the models under different lighting and rotation conditions. Both processes are described in detail in [19].

In short, the 3D faces were generated using the commercial software FaceGen^{TM}Modeller (`https://facegen.com`), which allows creating and exporting realistic facial models with different ages, genders, races, and facial expressions, in particular, 16 US-English visemes. We first used FaceGen to create and export 100 subjects for the training dataset. Then, we combined the models – that share the same topology, *i.e.*, an equal number of vertices and faces, and a full correspondence between every point – to produce new subjects via linear combinations. Applying the combinations (both on meshes and texture data) resulted in 2550 subjects.

Then, the C++ open source engine OGRE 3D (`http://www.ogre3d.org/`) was used for rendering the created models and grouping the images according to

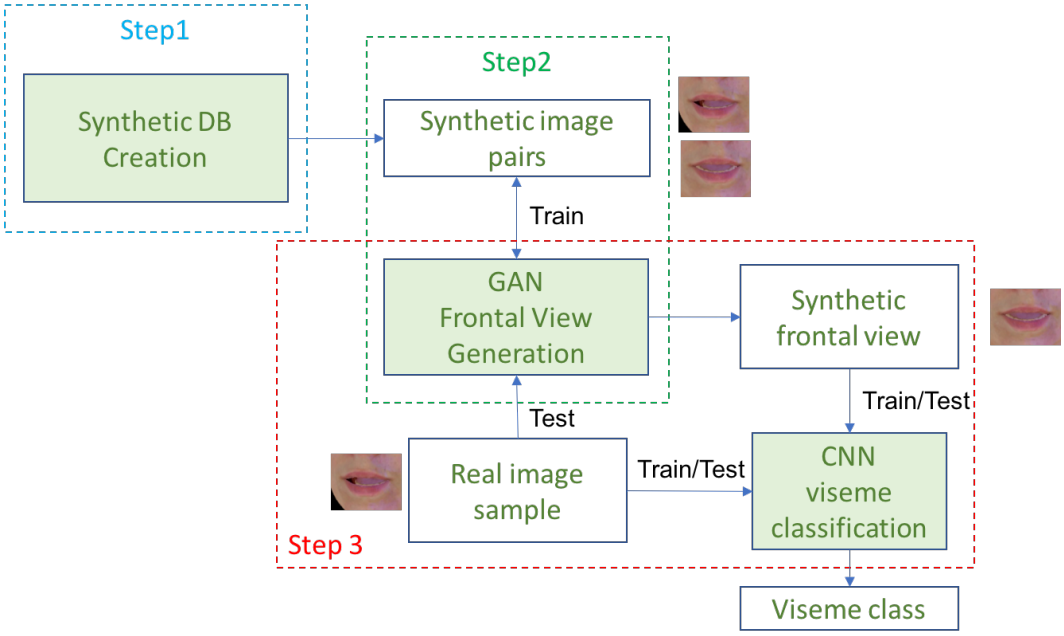


Figure 1. Our proposed methodology: first we generate a comprehensive synthetic viseme dataset; then we train a GAN to generate frontal views from a given random view image; and finally, we use both the given image and the corresponding GAN-generated frontal view image to classify the viseme.

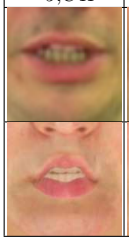
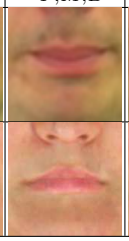
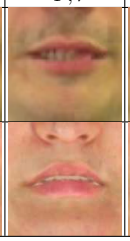
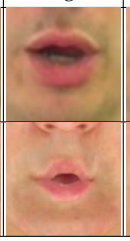
Consonants						Vowels				
$V_{J,CH}$	$V_{P,M,B}$	$V_{F,V}$	$V_{D,T,S}$	$V_{R,W}$	$V_{G,K,N}$	V_A	V_E	V_I	V_O	V_U
										

Table 1. Samples of visemes from each class. The first row of images shows the corresponding viseme displayed by a subject from the GRID dataset. The second row shows the same viseme displayed by a synthetic subject in the frontal view.

the fixed view angles. Using the engine, all subjects are rendered, one at a time, displaying each viseme. While the camera is fixed, and targeted at the face region, the subjects are rendered displaying each mouth shape under various lighting conditions. In each iteration, a face is loaded and four screenshots are taken under the following rotation conditions. First, random values are computed for defining the face rotation around x, y, and z axis. Then, the face is rotated (maintaining the same viseme and lighting condition) around the y axis in 0° (frontal view). For the random rotation, the angles are limited to [-30°,30°] for x axis, [-15°,15°] for y axis, and [-15°,15°] for z axis. The rotation in the y axis is small because the GRID dataset contains faces acquired in a near-frontal position.

This process is repeated while all subjects display every viseme. At the end, we have a dataset of images with the frontal view fixed angle paired with the images captured at random rotations. Each subject is rendered once displaying each viseme. In total, we generated a dataset with 2550 subjects and 16 visemes (including the neutral face for silence), totaling 40.800 synthetic images. A few samples are displayed in Figure 2.

3.3. GRID Dataset Annotation

The GRID dataset is composed of registered video and audio data recorded from 33 subjects in a controlled environment. Each pair of video and audio accompanies the corresponding transcription stating when the pronunciation of each word occurs. However, this annotation is not detailed in the phoneme level.

Therefore, we needed to use forced phonetic align-

Consonants			Vowels		
Viseme	Phoneme	Sound	Vis.	Phon.	Sound
$V_{J,CH}$	/ʤ/	**j**eep	V_A	/ɑː/	c**a**r
	/ʧ/	**ch**eap		/aʊ/	h**ou**se
	/ʃ/	dilu**t**ion		/aɪ/	fl**y**
	/ʒ/	delu**s**ion		/ʌ/	c**u**p
$V_{P,M,B}$	/p/	**p**it	V_E	/e/	**e**gg
	/b/	**b**it		/eɪ/	s**a**me
	/m/	**m**ap		/æ/	c**a**t
$V_{F,V}$	/f/	**f**at	V_I	/iː/	sh**ee**p
	/v/	**v**at		/ɪ/	sh**i**p
$V_{D,T,S}$	/d/	**d**in	V_O	/ɔː/	d**oo**r
	/t/	**t**in		/ɔɪ/	c**oi**n
	/s/	**s**ap		/əʊ/	n**o**se
	/z/	**z**ap			
	/θ/	**th**igh			
	/ð/	**th**y			
$V_{R,W}$	/r/	**r**un	V_U	/ʊ/	b**oo**k
	/w/	**w**e		/uː/	b**oo**t
$V_{G,K,N}$	/g/	**g**ut			
	/k/	**c**ut			
	/n/	thi**n**			
	/ŋ/	thi**ng**			
	/l/	**l**eft			
	/y/	**y**es			
	/h/	**h**am			

Table 2. Phoneme-to-viseme map used in this work. Notice that each viseme class groups a set of phonemes, exemplified here by one sound that they represent.

(a) Random view.

(b) Frontal view (0°).

Figure 2. Samples of the synthetic dataset. The random images (a) are paired to the corresponding image in the frontal view – (b)

ment[1] to estimate the location of each phoneme in the audio file, so we could extract the viseme at the corresponding frame from the video file. For this task, we used Prosody-lab [11], a python-based open source tool based on HTK (Hidden Markov Model Toolkit - `http://htk.eng.cam.ac.uk/`). Although Prosody-lab contains its own dictionary and acoustic model, using the audio and transcriptions from GRID for training a new acoustic model has proved to generate a more precise alignment.

Figure 3 displays the result of the phonetic alignment for the phrase "BIN BLUE AT F TWO NOW", visualized in the free software Praat (`www.praat.org`). More details on the acquisition of the real viseme dataset are presented in [20].

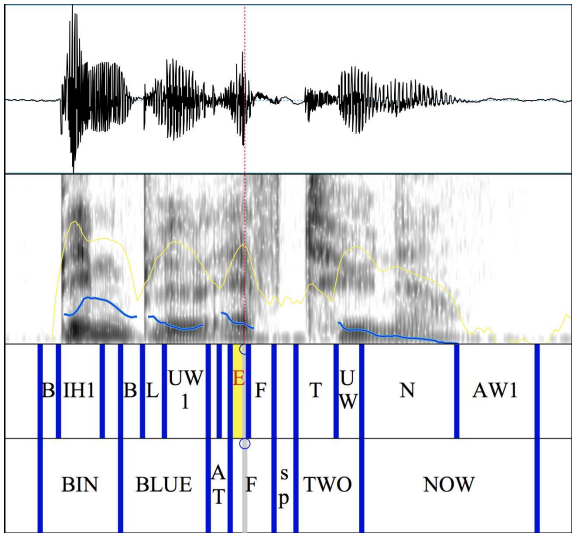


Figure 3. Alignment of words and phonemes (specified in the blue delimited intervals) of an audio recording from GRID. In the first row one observe the waveform, in the second the spectrogram, in the third and forth the time occurrences of spoken phonemes and words respectively. Over the spectrogram we observe the fundamental frequency contour in yellow and the energy contours in blue.

3.4. Frontal view GAN-based Mapping

To map from random views to frontal views we used the Generative adversarial networks firstly proposed by Goodfellow *et al.* [10] and widely discussed in the Computer Vision community in the last few years. They are basically composed by two networks: a *generator* that outputs invented images; and a *discriminator* that evaluates them against real ones. In this schema, optimization is dual: the generator tries to generate images realistic enough to fool the discriminator, that tries to correctly identify the invented images as non-real, as shown in Figure 4. Each of them is optimized in rounds, and in practice one expects that the generative network learns to map from a latent space to a particular data distribution of interest, and the discriminator network learns to discriminate between instances from the true data distribution and candidates produced by the generator. Since the objective of the generator network is to increase the error rate of the discriminator network, very realistic samples are expected as an outcome from the generator network.

[1]Forced phonetic alignment is the process of determining the times at which individual sounds appear in an audio recording – under the constraint that words in the recording follow the same order as they appear in an accompanying transcript file.

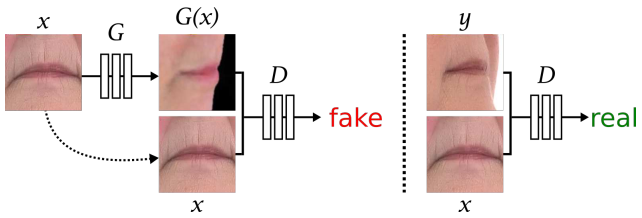


Figure 4. The idea behind Pix2Pix. The generator G receives an image x with view taken at a random angle, and tries to generate an image $G(x)$ with a view at a fixed angle. The discriminator, D, learns to classify between fake (synthesized by the generator) and real pairs of images.

In this work, we used specifically the Pix2Pix GAN architecture proposed by Isola *et al.* [13], which is composed by a classic encoder-decoder generator network and a "PatchGAN" discriminator network. The Pix2Pix generator network is composed of several encoder convolutional blocks, followed by several decoder blocks. Encoder blocks are convolutional networks that receive an input image and reduce it to a much smaller data representation trough convolutions, while decoder blocks do the opposite and try to recover the original information (or any different goal information) out of this compact data representation. The discriminator network consists of a sequence of encoders where the last block outputs a representation where each pixel encodes how believable the corresponding image patch is with respect to a ground truth image pair.

The training of such an architecture consists of two steps:

1. Training the discriminator network using data generated by the generator network and real data, so it learns to discriminate between a real and an invented pair of images;

2. Training the generator network using the pair of invented and reference images and the discriminator guess as a bias to update the generator gradients.

In our experiments, we used the synthetic dataset described before as training data to a Pix2Pix network, with the goal of creating a viseme face image at a fixed frontal view given an image captured at a random angle. In this implementation, the generator gets an image with a view from a random angle and tries to generate the image at a frontal view, and the discriminator tries to identify if the generated image is at the right angle or not. With the trained GAN, we were able to feed the network with real images and get their corresponding synthetic frontal view.

3.5. CNN viseme classification

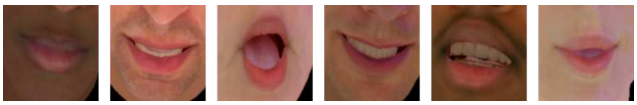
With a pair of real image and the corresponding GAN-generated frontal view, we trained a CNN to classify visemes. In this work, we used the Xception architecture proposed by [7] for classification, which is one of the top ranked CNNs for multi-class categorical image classification. The network is inspired in the classic Inception architecture, where Inception modules are replaced with faster depth-wise separable convolutions.

4. Experiments

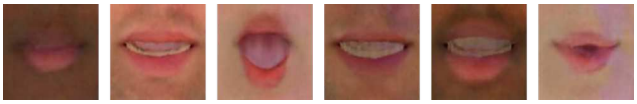
Our experiments consisted of training the GAN to generate the synthetic frontal view, and training the CNN to classify visemes.

4.1. Synthetic frontal view generation

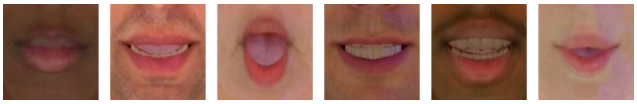
The goal of this experiment was to assess the visual similarity between the images generated by the GANs and a ground truth. The ground truth was created by rotating the actual 3D model to the frontal view, and 20k images where used as test set for qualitative evaluation. Some results of this experiment are displayed in Figure 5. Notice that the images outputted by the GAN are very similar to the ground truth (target) images, except for some minor artifacts that occur especially at the teeth region.



(a) Input images at random view.



(b) Output of GAN when mapping images in (a) to the frontal view.



(c) Ground truth of images in (a) at frontal view.

Figure 5. Results when mapping synthetic inputs to the frontal view.

We also visually inspected the synthetic frontal view generated using real GRID data for different visemes. The visual outcome is displayed in Figure 6. One can notice that any rotation effect is corrected, and the output image looks like the synthetic ones, which we also believe to be helpful for the classification step. Again, some minor artifacts are observed especially at the teeth region.

4.2. Viseme classification

The classification experiments were planned to evaluate the benefits of using the synthetic frontal view to improve the performance of CNNs on the classification of visemes from real data. They consisted of two different rounds: (i) using only GRID dataset; and (ii) using GRID im-

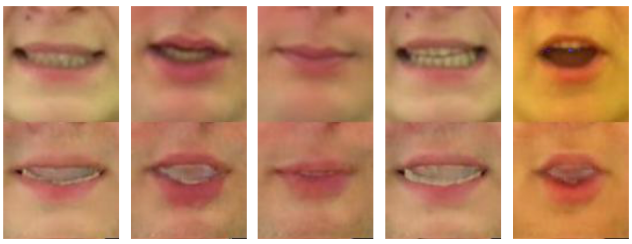


Figure 6. Results when mapping real data inputs to the frontal view for different visemes. The first and second rows show respectively real and corresponding synthesized data for different visemes.

age concatenated with the GAN-generated synthetic frontal view, as shown in Figure 6.

The GRID dataset used for training comprised 29 subjects and generated an imbalanced training set of viseme images, ranging from 26788 samples of class $V_{D,T,S}$ to 868 samples of class V_O. Validation and testing used 2 subjects each and generated an image set ranging from 2145 samples of class $V_{D,T,S}$ to 75 samples of class V_O for test, and from 1391 samples of class $V_{D,T,S}$ to 63 samples of class V_O for validation.

All the experiments considered the Xception architecture proposed by [7], Adam algorithm was used for weights optimization and categorical cross entropy as the loss function. Training was run over 30 epochs (which was a safe margin to achieve convergence in our experiments), where each epoch consisted of 500 balanced batches of 55 samples – 5 samples per viseme class, taken randomly from the training set. Validation consisted of 50 steps of 55 samples – again, 5 samples per viseme class, taken randomly from the validation set. Testing considered all image samples from the test set up to the limit of 1000 samples per class.

Our two experiments used exactly the same CNN architecture and basal data, the difference is that the first run used only GRID original data trained from scratch, and the second run used an image composed by the concatenation of a real image and the corresponding GAN-generated frontal view, also trained from scratch. The first model achieved an overall accuracy on testing of 61.40%, while our proposed methodology achieved 67.30% overall accuracy, which represents an increase of **5.9%** in accuracy. Detailed results are presented by means of a confusion matrix displayed in Table 3. It is possible to observe that our model outperformed the baseline model in at least 4 visemes by a great margin, under performed in 2 visemes, and got similar results in 5 visemes. It is also possible to observe that the differences in accuracy are especially significant for visemes $V_{F,V}$, V_I and V_U.

In comparison to other works, Saenko *et al.* [21] used SVM to achieve 63% accuracy on distinguishing between 6 viseme classes and Koller *et al.* [14] used deep CNNs to obtain 55.7% accuracy on the recognition of 12 viseme classes. We also achieved an accuracy 2.5% superior in comparison with the transfer learning method presented in [20], which achieved 64.80% of overall accuracy using also GRID database and 11 visemes.

5. Conclusion

This work aimed at improving viseme classification using a novel GAN-based solution for easing the CNN-based classification task. Our results indicate that using GANs to generate a perfectly locked frontal view of a given input mouth shape was able to improve in **5.9%** the accuracy of CNN-based viseme recognition. Given the challenging nature of this task, demonstrated by our literature review, this improvement is considered to be significant. We also observed that the GAN model used was able to generate close-to-real synthetic frontal views from real images and diminish observed rotation effects.

An advantage of using synthetically generated databases for training models for further transfer learning in any creative way, is to be able to control the dataset conditions. In this application we used random near frontal views, which were observed in the GRID dataset. However, it is worth mentioning that such methodology could be applied to different conditions, and this is a natural extension of this work.

For further research, we intend to explore the use of GANs for generating synthetic multi-view images in the context of CNN-based viseme classification, and try different CNN architectures for visemes classification. As related research, since our 3D models may display combined expressions, we intend to explore emotions recognition and combine visemes and facial expressions for evaluating visemes recognition in videos where the speech is being affected by different emotions, such as happiness or anger.

References

[1] I. Abbasnejad, S. Sridharan, D. Nguyen, S. Denman, C. Fookes, and S. Lucey. Using synthetic data to improve facial expression analysis with 3d convolutional networks. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 1609–1618, Oct 2017.

[2] A. H. Abdelaziz. Improving acoustic modeling using audio-visual speech. In *2017 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1081–1086, July 2017.

[3] Amazon Polly. Phoneme/viseme tables for supported languages. https://docs.aws.amazon.com/polly/latest/dg/ref-phoneme-tables-shell.html. Accessed: 2018-02.

[4] Y. M. Assael, B. Shillingford, S. Whiteson, and N. de Freitas. Lipnet: Sentence-level lipreading. *CoRR*, abs/1611.01599, 2016.

[5] H. L. Bear and R. Harvey. Phoneme-to-viseme mappings: the good, the bad, and the ugly. *Speech Communication*, 95(Supplement C):40 – 67, 2017.

GRID visemes classification: using synthetic data weights initialization and training from scratch (in parenthesis)											
	$V_{J,CH}$	$V_{P,M,B}$	$V_{F,V}$	$V_{D,T,S}$	$V_{R,W}$	$V_{G,K,N}$	V_A	V_E	V_I	V_O	V_U
$V_{J,CH}$	**728** (609)	40 (62)	1 (0)	0 (0)	14 (12)	1 (4)	0 (0)	200 (311)	0 (0)	0 (1)	17 (2)
$V_{P,M,B}$	12 (4)	959 **(969)**	1 (0)	0 (0)	0 (0)	1 (0)	0 (0)	28 (28)	0 (0)	0 (0)	0 (0)
$V_{F,V}$	8 (0)	0 (1)	**498** (233)	0 (2)	12 (2)	84 (102)	47 (24)	68 (81)	26 (6)	205 (549)	53 (1)
$V_{D,T,S}$	0 (0)	1 (1)	1 (0)	98 **(167)**	104 (15)	0 (0)	1 (2)	0 (0)	2 (3)	7 (38)	12 (0)
$V_{R,W}$	5 (35)	1 (1)	0 (0)	19 (96)	681 **(769)**	1 (0)	0 (0)	73 (70)	13 (10)	1 (11)	207 (9)
$V_{G,K,N}$	0 (0)	0 (0)	17 (8)	0 (0)	2 (1)	**876** (867)	3 (2)	7 (0)	0 (0)	3 (33)	3 (0)
V_A	0 (0)	0 (0)	0 (3)	0 (0)	0 (0)	0 (0)	53 **(54)**	0 (0)	0 (0)	0 (2)	19 (13)
V_E	84 (57)	20 (19)	61 (24)	1 (8)	16 (3)	14 (9)	50 (25)	401 **(464)**	9 (6)	43 (123)	40 (1)
V_I	0 (0)	5 (5)	10 (1)	110 (86)	225 (133)	36 (26)	1 (0)	15 (48)	**427** (85)	168 (617)	4 (0)
V_O	0 (0)	0 (0)	68 (35)	11 (23)	34 (2)	1 (1)	20 (14)	4 (2)	18 (0)	204 **(309)**	33 (7)
V_U	0 (17)	0 (0)	0 (0)	0 (0)	2 (17)	0 (0)	3 (10)	0 (0)	0 (0)	0 (0)	**70** (31)

Table 3. Confusion matrices for GRID dataset. The first row in each cell shows the results of the CNN trained using real data concatenated with GAN-generated frontal view. The second row shows the results of the CNN trained only using real data.

[6] H. L. Bear, R. Harvey, B. J. Theobald, and Y. Lan. Resolution limits on visual speech recognition. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 1371–1375, 2014.

[7] F. Chollet. Xception: Deep learning with depthwise separable convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1251–1258, 2017.

[8] J. S. Chung, A. W. Senior, O. Vinyals, and A. Zisserman. Lip reading sentences in the wild. *CoRR*, abs/1611.05358, 2016.

[9] M. Cooke, J. Barker, S. C., and X. Shao. An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, 120(5):2421–2424, 2006.

[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014.

[11] K. Gorman, J. Howell, and M. Wagner. Prosodylab-aligner: A tool for forced alignment of laboratory speech. *Canadian Acoustics*, 39(3):192–193, 2011.

[12] 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)*, pages 3574–3582, 2016.

[13] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. *arxiv*, 2016.

[14] O. Koller, H. Ney, and R. Bowden. Deep learning of mouth shapes for sign language. In *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, pages 477–483, 2015.

[15] A. Koumparoulis, G. Potamianos, Y. Mroueh, and S. J. Rennie. Exploring ROI size in deep learning based lipreading. In *14th International Conference on Auditory-Visual Speech Processing*, 2017.

[16] Y. Lan, R. Harvey, B. Theobald, E. Ong, and R. Bowden. Comparing visual features for lipreading. In *International Conference on Auditory-Visual Speech Processing*, pages 102–106, 2009.

[17] S. Lee and D. Yook. *Audio-to-Visual Conversion Using Hidden Markov Models*, pages 563–570. Springer Berlin Heidelberg, Berlin, Heidelberg, 2002.

[18] I. Matthews, T. F. Cootes, J. A. Bangham, S. Cox, and R. Harvey. Extraction of visual features for lipreading. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(2):198–213, 2002.

[19] A. B. Mattos and D. A. B. Oliveira. Multi-view mouth renderization for assisting lip-reading. In *Proceedings of the 15th Web for All Conference*, W4A '18, 2018.

[20] A. B. Mattos, D. A. B. Oliveira, and E. da Silva Morais. Improving CNN-based viseme recognition using synthetic data.

In *2018 IEEE International Conference on Multimedia and Expo*, 2018.

[21] K. Saenko, K. Livescu, M. Siracusa, K. Wilson, J. Glass, and T. Darrell. Visual speech recognition with loosely synchronized feature streams. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 2, pages 1424–1431 Vol. 2, 2005.

[22] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman. Synthesizing obama: Learning lip sync from audio. *ACM Trans. Graph.*, 36(4):95:1–95:13, July 2017.

[23] K. Thangthai, H. L. Bear, and R. Harvey. Comparing phonemes and visemes with dnn-based lipreading. In *Proceedings of British Machine Vision Conference*, page In Press. BMVA Press, 2017. © 2017 The authors.

[24] G. Z. X. H. M. P. Z. Zhou. A review of recent advances in visual speech decoding. *Image and Vision Computing*, 32(9):590 – 605, 2014.