Proc. of the International Conference on Electrical, Computer and Energy Technologies (ICECET 2022)
20-22 July 2022, Prague-Czech Republic

GAN-Based Augmentation for Gender Classification from Speech Spectrograms

Hynek Bořil
ECE Department
University of Wisconsin–Platteville
Platteville, WI, U.S.A.
borilh@uwplatt.edu

Skyler Horn
ECE Department
University of Wisconsin–Platteville
Platteville, WI, U.S.A.
skylerhorn8@gmail.com

***Abstract*—The focus of this study is on gender classification from speech signals produced by adults. Automatic estimation of gender has a broad variety of applications ranging from forensics, authentication systems, diarization of meetings, or user-centered interactive agents, and plays a crucial role also in 'internal' technological solutions aimed at improving model accuracy, such as selection of gender-specific acoustic models. In this study, we explore scenarios where only a limited amount of real training data is made available for training of a 2-D convolutional neural network classifier. To address sparsity of the training data, the training set is augmented by synthetically generated samples produced by a generative adversarial network. The adversarial network is given access only to the same limited real-world training dataset as the gender classifier. We demonstrate that even when 80 % of the already limited training data are removed and replaced by synthetic spectrograms, the gender classifier models can still be successfully trained thanks to the augmentation and maintain competitive performance.**

***Index Terms*—gender detection from speech, GAN-based augmentation, synthetic spectrograms**

I. Introduction

Spoken language represents one of the most prominent forms of human communication and, with the recent advancements in speech technologies, starts to become a valid modality also for human–computer interactions. Acoustic speech signals contain both linguistic information, which can be transcribed to text, and paralinguistic information about the speaker's identity [1], gender [2], speaking style [3], emotional state and cognitive load [4], stress [5], physiological properties (e.g., height [6] and age [2]), or health condition [7]. In this study, our focus is on gender estimation from adult speech. Gender information is instrumental for numerous applications in forensics, security, personalized spoken dialog systems and automated interactive agents, automatic segmentation/diarization of audio streams, and can be utilized during training and deployment of speech based systems for speech feature equalization or building and engaging gender-specific acoustic models for better performance [8]. Knowledge of a subject's gender can help improve modeling and prediction of other paralinguistic speaker traits from speech [9].

Gender traits are displayed in speech signals across various feature dimensions. Following the linear model of speech production (see Fig. 1, [10]), parameters such as fundamental frequency F_0 [11], [12], vocal quality and pitch [13], spectral slope [11], vocal tract length [14], and acoustic-phonetic spaces [11], [15] have been known to be impacted by gender of the speaker.

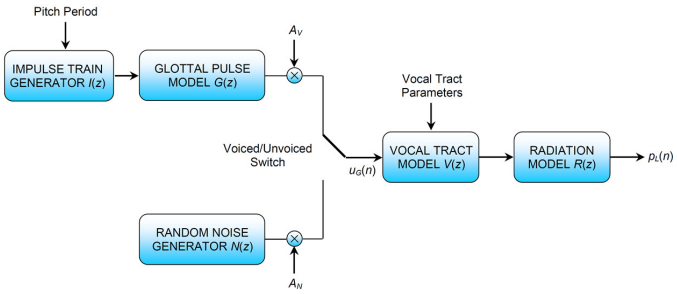

Fig. 1. Linear model of speech production.

Initial work on automatic gender recognition typically employed handcrafted features and a variety of classification back-ends. [16] used speech formant estimates as features for gender classification. [17] trained two hidden Markov models (HMM) to represent statistical distributions of pitch in males and females while [18] combined acoustic and prosodic features. The authors in [19] performed gender classification on short-term spectra which were extracted from 1-second segments and passed to a neural network classifier. Later, [20] proposed a two-stage classifier where the first stage identified gender based on pitch and samples identified as ambiguous were passed to the second stage that utilized a Gaussian Mixture Model (GMM) classifier. [21] combined acoustic, prosodic, and voice quality features and fused multiple back-end classifiers while [22] adopted paradigms popular from speaker identification, GMM–Universal Background Model (GMM–UBM) and GMM–Support Vector Machines (GMM–SVM), together with a Mel Frequency Cepstral Coefficients (MFCC) front-end.

978-1-6654-7087-2/22/$31.00

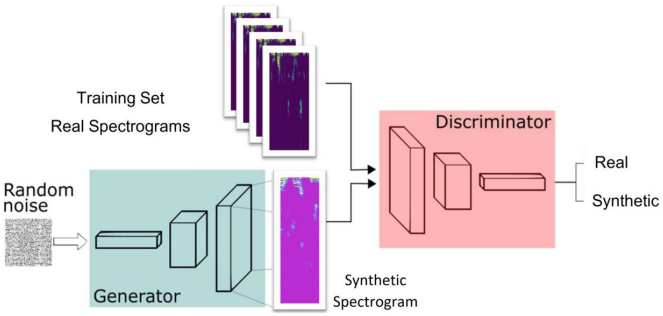

Fig. 2. Learning and generation of GAN-based spectrograms.

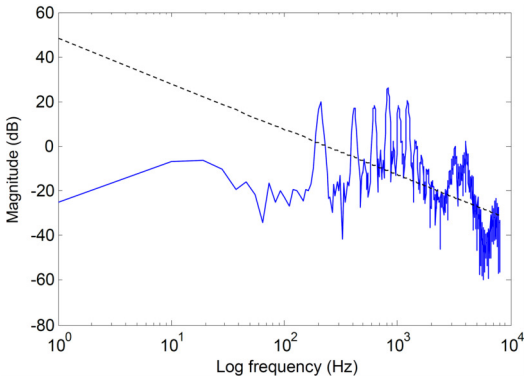

Fig. 3. Example of spectral slope extraction from short-term spectrum.

The emergence of deep learning over a decade ago, followed by discoveries of new neural network architectures, have enabled tackling speech engineering tasks that were previously confined to laboratory settings. At the same time, due to the extensive number of trainable parameters in deep learning models, modern speech systems typically require access to larger amounts of training data compared to previous machine learning solutions. Many recent gender classifiers rely on deep learning and are typically implemented as end-to-end systems; for example [2] applied convolutional neural network (CNN) with a spatially designed multi-attention module (MAM) on speech spectrograms. This being said, other state-of-the-art studies successfully combine handcrafted features and deep learning strategies, such as [23] where a 1-D CNN was used to model MFCC, Mel spectrogram, and Chroma features.

Our focus in this study is on speech-based gender classification in conditions where only sparse data are available for training. We build a 2-D CNN-based gender classifier derived from the AlexNet architecture [24] and train it on a very small subset of the LibriSpeech database [25]. In order to aid the training process, we augment the available real samples by synthetic samples generated by a Generative Adversarial Network (GAN) [26]. This GAN is trained on the same small set of available speech samples and its task is to generate gender-specific speech spectrograms that would meaningfully reflect gender-specific traits seen in real spectrograms. These artificial spectrograms are then used to augment the real speech sample spectrograms for the CNN gender classifier training. We demonstrate that with such augmentation, even if the amount of real training samples is reduced to about 20 % of the original size, the GAN-based augmentation assures nearly intact classification performance. We disclose that a preliminary version of this study was presented in the form of a 1-paragraph abstract in [27].

The remainder of the text is organized as follows. Section 2 provides a brief overview of regularization techniques used in deep learning. Section 3 outlines the experimental setup, analyzes selected speech production features known to carry gender traits, presents examples of GAN-generated spectrograms, and details results of gender classification experiments under various augmentation conditions and in the presence of additive noise. Section 4 concludes the study.

II. Regularization in Deep Learning

Deep learning has been successfully applied to various domains including image processing, speech recognition, or medical imaging. Data coming from these domains pose similar challenges when it gets to modeling and classification–high variability of feature distributions, limited access to data, and limited/costly access to ground truth transcriptions (labels). Given the large number of trainable parameters in a typical deep learning model, there is a high risk of overtraining in scenarios where limited training data are available.

One approach to improve generalization performance is through modification of the model's architecture. For example, modified architectures stemming from AlexNet have been recently introduced–VGG-16 [28], ResNet [29], Inception-V3 [30], and DenseNet [31].

Another way to reducing the danger of overtraning is regularization. The purpose of regularization techniques is to assure good generalization and robustness of the models [32], [33], [34], [35]. Some approaches to regularization are adding a term into the loss function [32], dropout regularization that masks activation values of random neurons during training [36], batch normalization [37], transfer learning [38], pretraining [39], and one-shot and zero-shot learning [40]. Data augmentation addresses the issue of overfitting by introducing a larger, more comprehensive dataset. There are numerous data augmentation techniques available, all relying on the assumption that the original data can provide additional information to the training process if they are processed by various transformations. Such transformations are often introduced via data warping or oversampling [35]. Warping can transform existing samples via geometric [41] and color [42] manipulations, random erasing [43] and noise adding [44] while oversampling creates synthetic samples that are used to extend the existing dataset.

Generative Adversarial Networks (GANs) [26] provide an excellent means for synthetic image generation [45] and have been employed in image synthesis from text [46], generation of high-resolution images from low-resolution sources [47], image-to-image translation [48], high-resolution image blending [49], restoration of missing parts in images [50],

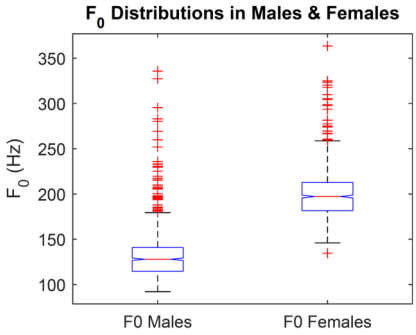

Fig. 4. Fundamental frequency F_0 distributions in male and female subjects.

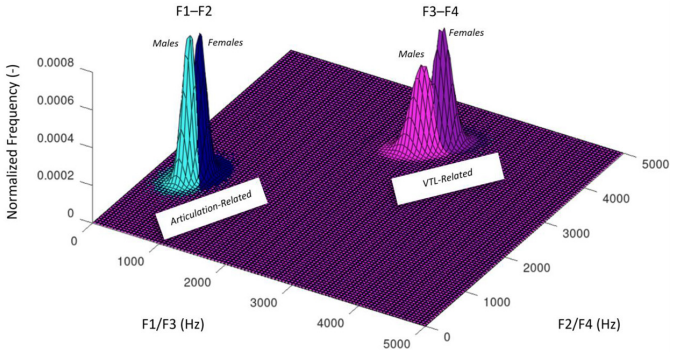

Fig. 6. F_1–F_2 and F_3–F_4 formant distributions; voiced segments.

image denoising, and brain MRI synthesis [51]. The GAN structure is outlined in Fig. 2 and can be broken down into two parts–(i) a generator which is trained to produce realistic images and (ii) a discriminator which is trained to distinguish authentic vs. synthetic images. Both the generator and discriminator are trained together. The generator's goal is to fool the discriminator while the discriminator gets better and better in distinguishing real and synthetic images, which in turn forces the generator to refine its outputs to better match real-world references. Once the GAN training is completed, the output from the generator is used to produce synthetic images and the generator is usually discarded. As indicated in the literature review above, GAN's are excellent candidates for oversampling-style data augmentation and have been successfully employed in that function in various domains. Our goal in this study is to investigate efficiency of GAN-based augmentation in the context of gender classification from speech spectrograms.

III. Experiments

The following sections outline results of speech production analyses, present examples of GAN-generated spectrograms, and study performance of a CNN-based gender classifier when given access to various augmentation rates during training and when exposed to speech samples corrupted by additive noise at various signal-to-noise ratios (SNR).

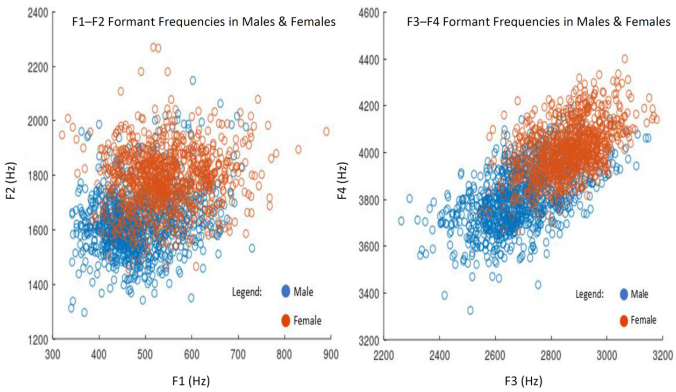

Fig. 5. Formant F_1–F_2 and F_3–F_4 scatter plots for male and female subjects; voiced segments.

A. Material

The speech samples utilized in the following experiments are drawn from the LibriSpeech corpus [25]. The training set consists of twenty male and twenty female sessions, the cross-validation set is formed by ten male and females sessions, each, and the open test set also consists of ten male and female sessions, each. Each speaker session comprises 23 utterances. The training, cross-validation, and test sets contain mutually exclusive speakers (no speaker overlap). When all samples available in all the speaker session are preserved, the total duration per each set is as follows: *Male Training*–63 minutes, *Female Training*–58 minutes, *Male CV*–29 minutes, *Female CV*–29 minutes, *Male Test*–30 minutes, and *Female Test*–25 minutes. For the purpose of the following speech production analyses, all speaker session samples were preserved and the training, CV, and test sets were pooled together per gender. For the GAN and CNN classifier training, only a limited subset of the samples is made available.

B. Analysis of Gender Traits in Speech Production

As outlined in Introduction, a number of speech production parameters can be expected to carry gender cues. To verify the rate of these acoustic cues in the subset of LibriSpeech,

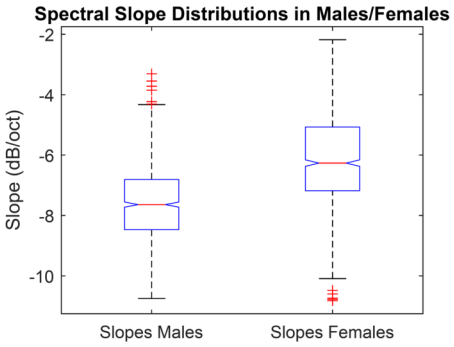

Fig. 7. Spectral slope distributions in male and female subjects.

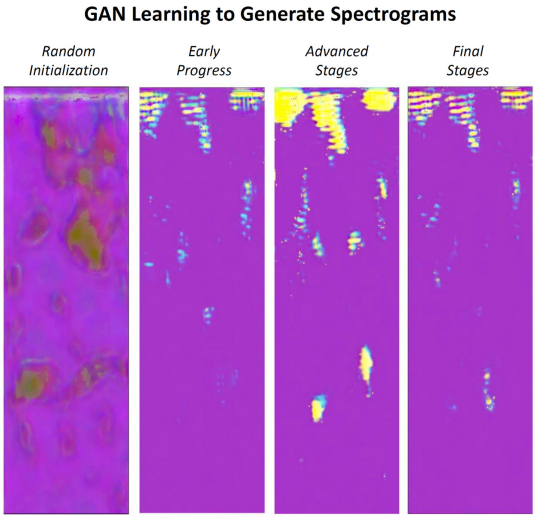

Fig. 8. Example of GAN learning stages in spectrogram generation.

fundamental frequency F_0, first four formant center frequencies (F_1–F_4) and spectral slopes are analyzed. Fundamental frequency and formant frequencies were extracted using WaveSurfer [52]. Spectral slopes were extracted by a script written in Matlab which fits straight lines (by means of linear regression) into short-term logarithmic spectra plotted over a logarithmic frequency axis (see an example in Fig. 3). Formant frequencies and spectral slopes were extracted only from voiced segments of the utterances, where F_0 labels were used to identify voiced island boundaries. Fig. 4 provides box plots for male and female F_0 distributions.

Fig. 5 shows F_1–F_2 and F_3–F_4 formant scatter plots for both genders. It can be seen that while the acoustic-phonetic spaces overlap in the formant planes, female formant center frequencies tend to be higher across all four formants. Formant distributions are further detailed in Fig. 6. Finally, spectral slope box plots are shown in Fig. 7. It can be seen that for all studied production parameters–fundamental frequency, first four formant locations, and spectral slopes in voiced segments, there are prominent differences between genders.

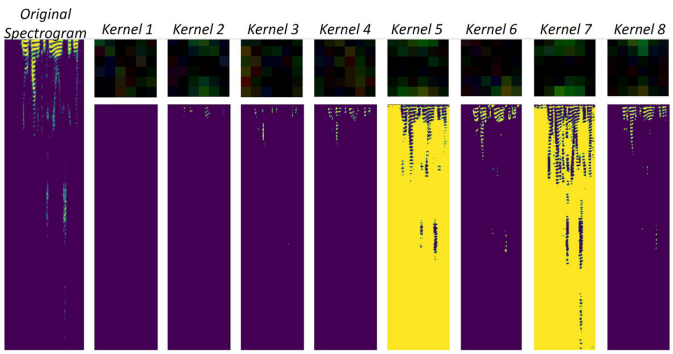

Fig. 9. Gender classifier–learned convolutional layer kernels and examples of kernel-filtered spectrograms.

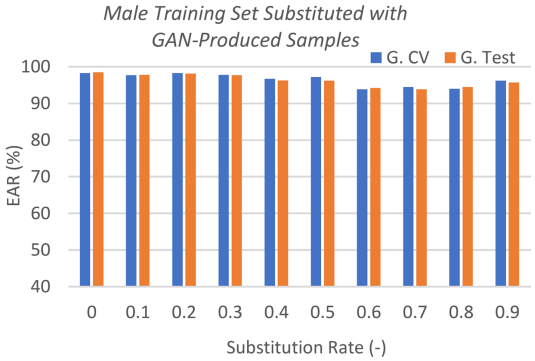

Fig. 10. Effect on substitution of authentic training samples by GAN-generated spectrograms on gender classification performance; male training data substitutions; training on 100 samples per gender; *CV*–performance on cross-validation set, *Test*–performance on open test set.

C. GAN-Based Spectrogram Generation

This section discusses experiments with GAN-based synthetic spectrogram generation. As outlined in Sec. 2, our GAN contains generator and discriminator modules. The generator consists of a dense layer, reshape, 3 x bilinear 2-D upsample, and convolutional 2-D transpose layers and outputs 3 x 400 x 160 pixel bitmaps (where 3 indicates the 3 R, G, B components). Our discriminator block consists of two discriminator modules. Discriminator 1 operates on full R, G, B bitmaps and contains one convolutional layer with twelve 4 x 4 kernels followed by a dense layer with 16 neurons. Discriminator 2 applies 4 x 4 max pooling on the R, G, B bitmaps, followed by a convolutional layer with six 4 x 4 kernels and a dense layer with 12 neurons. Fig. 8 shows an example of spectrogram bitmaps generated by the GAN shortly after its random initialization and at subsequent stages of its training. It can be see that as the training progresses, the generated bitmaps gain stronger resemblance with real speech spectrograms. At the same time, the example demonstrates that even after the GAN training is fully completed, the generated

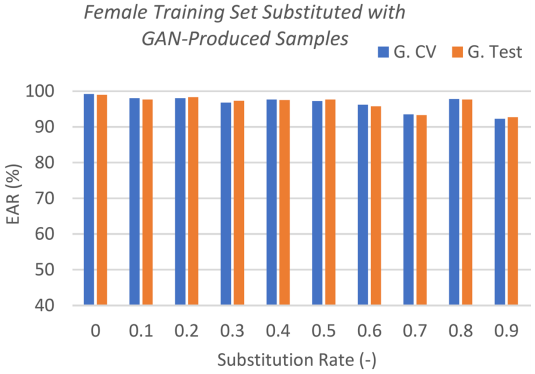

Fig. 11. Effect on substitution of authentic training samples by GAN-generated spectrograms on gender classification performance; female training data substitutions; training on 100 samples per gender; *CV*–performance on cross-validation set, *Test*–performance on open test set.

synthetic spectrograms contain components that may reveal their artificial nature to a trained eye. This being said, as will be shown in the following section, these artifacts do not seem to have a prominent impact on the augmentation process and success of gender classifier training.

D. Gender Classification with Augmented Training

The architecture of our CNN-based gender classifier is derived from AlexNet [24]. The input convolutional layer has 400 x 160 x 3 neurons and implements eight 8 x 8 2-D kernels, followed by max pooling, flattening, a dense layer with 32 neurons, and an output producing gender scores. For the classification experiments, a subset totalling at 100 samples (utterances) per gender was used as the training set. These samples were uniformly drawn from the twenty male and twenty female speaker sessions outlined in Sec. III.A. In the baseline experiment, When retraining the CNN model from scratch on this training set, the equal accuracy rate (EAR), which is calculated as a complement of equal error rate (EER), $EAR = 100 - EER$ (%), was averaging at 99 %, which indicates that this training set is sufficient for training without any need for augmentation. Kernels learned by the CNN layer are shown in Fig. 9, together with an example of an input spectrogram and its filtered versions by the kernels. It can be seen that each kernel highlights different regions in the spectrogram that are then passed to the following layers. Given the high accuracy of the classifier, the highlighted regions can be expected to carry strong gender-specific information.

In the next step, the training set was reduced to 50 samples per gender, causing the EAR to drop to 67.4 %. This indicates that the model does not have access to sufficient amount of data to avoid overfitting and hence, data augmentation might be necessary. To explore the impact of GAN-based augmentation, we iterated over varying sizes of real training data that were made available for training of the CNN classifier from scratch. In the initial iteration, the first CNN classifier was given access to all 100 samples per gender for training. In the second iteration, the second CNN classifier had access to only 90 % of the real training samples for one gender and the missing 10 % were replaced by synthetic spectrograms generated by the GAN. At the same time, the GAN itself had access only to the same 90 % of the real samples for its training as the CNN classifier. Progressively, in each iteration of this experiment, further 10 % of the real samples were removed for the the selected gender's training set, both for the GAN and CNN models. At the same time, the GAN was tasked with supplying the CNN with increasing number of synthetic spectrograms so the total size of the training set (real + synthetic samples) would stay constant at 100 samples per gender. Results of this experiments are detailed in Fig. 10 and 11. It can be seen that there is some performance variability with different rates of real data substitution by GAN samples. However, this may not be necessarily only due to the data substitutions. Given that each GAN and CNN training starts with random initialization, neither the GAN generator or the CNN gender classifier are guaranteed to converge with the same success rate every time even if an identical training set is provided. However, in spite of these considerations, it can be seen that replacing up to about 80 % of the real samples by GAN-generated spectrograms helped maintaining a competitive classification accuracy which by far exceeded the 67.4 % EAR seen in the initial experiment with 50 % training data reduction and no augmentation engaged.

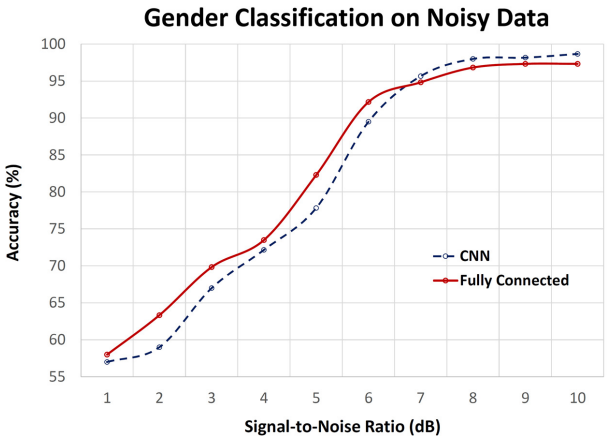

Fig. 12. Comparison of CNN-based and fully-connected gender classifier performance on noisy data; training on 100 samples per gender contaminated by white noise at SNR of 7 dB.

Finally, robustness of the CNN gender classifier was compared to a fully connected setup. Here, the 100 real training samples per gender were mixed with white Gaussian noise at 7 dB SNR and used for the CNN/fully connected (FC) network training. The CNN and FC gender classifiers were then exposed to the test set corrupted by white noise at SNR's from 1 to 10 dB SNR. As can be seen in Fig. 12, the FC setup provides somewhat better performance at low SNR's up to 6 dB. From 7 dB up, the CNN setup has a slight edge. It is also noted that from 8 dB up, both the CNN and FC setups operate at accuracy rates comparable to the clean test set scenario. This demonstrates benefits of noise-based data augmentation.

IV. Conclusions

This study investigated efficiency of GAN-based data augmentation in the context of gender classification from speech. The GAN generator was trained to produce gender specific synthetic spectrograms that were then used to extend a sparse real data training set. The evaluation experiments suggest that augmentation with synthetic spectrograms can notably improve success of a CNN model training when only limited real training samples are available. In addition, we have experimentally confirmed benefits of data augmentation via additive noise, where gender classifiers trained on noisy data were able to reach competitive performance from approximately 8 dB SNR. In overall, our study demonstrates that data augmentation strategies borrowed from the field of image processing can be successfully applied also in processing of 1-D acoustic signals in the context of gender classification.

REFERENCES

[1] J.H.L. Hansen and T. Hasan, "Speaker Recognition by Machines and Humans: A tutorial review," in IEEE Signal Processing Magazine, vol. 32, no. 6, pp. 74-99, Nov. 2015.

[2] A. Tursunov, M. Mustaqeem, J.Y. Choeh, S. Kwon, "Age and Gender Recognition Using a Convolutional Neural Network with a Specially Designed Multi-Attention Module through Speech Spectrograms," Sensors (Basel). 2021, 21(17):5892.

[3] H. Bořil, J.H.L. Hansen, "Unsupervised Equalization of Lombard Effect for Speech Recognition in Noisy Adverse Environments," in IEEE Transactions on Audio, Speech, and Language Processing, 18(6), 1379-1393.

[4] H. Bořil, O. Sadjadi, T. Kleinschmidt, J.H.L. Hansen, "Analysis and Detection of Cognitive Load and Frustration in Drivers' Speech," in Proc. of Interspeech'10, 502-505, Makuhari, Chiba, Japan, 2010.

[5] J.H.L Hansen, E. Ruzanski, H. Bořil, J. Meyerhoff, "TEO-Based Speaker Stress Assessment Using Hybrid Classification and Tracking Schemes," International Journal of Speech Technology, Springer, June 2012.

[6] J.H.L Hansen, K. Williams, K., H. Bořil, "Speaker height estimation from speech: fusing spectral regression and statistical acoustic models," Journal of the Acoustical Society of America (JASA), 138(2), Aug. 2015, 1052-1067.

[7] D. Braga, A.M. Madureira, L. Coelho, R. Ajith, "Automatic detection of Parkinson's disease based on acoustic analysis of speech," Engineering Applications of Artificial Intelligence, Volume 77, 2019, Pages 148-158,

[8] S. Wegmann, D. McAllaster, J. Orloff, and B. Peskin, "Speaker normalization on conversational telephone speech," in Proceedings of ICASSP 1996, vol. 1, pp. 339-341, 1996.

[9] S. I. Levitan et al., "Automatic identification of gender from speech," Speech prosody, 84-88, 2016.

[10] L. Rabiner, R. Schafer, "Digital Processing of Speech Signals," Englewood Cliffs: Prentice Hall, 1978.

[11] H. Bořil, "Robust speech recognition: Analysis and equalization of Lombard effect in Czech corpora,"Ph.D. dissertation, Czech Technical University in Prague, Czech Republic, 2008.

[12] J. Bishop, P. Keating, "Perception of pitch location within a speaker's range: Fundamental Frequency, voice quality and speaker sex," in The Journal of the Acoustical Society of America, vol. 132-2, pp. 1100-1112, 2012.

[13] C. G. Henton, "Fact and fiction in the description of male and female pitch," in Language and Communication, vol. 9, pp. 299-311, 1989.

[14] D.R. Smith and R.D. Patterson, "The interaction of glottal-pulse rate and vocal-tract length in judgments of speaker size, sex, and age," in The Journal of the Acoustical Society of America, vol. 118-5, pp. 3177-3186, 2005.

[15] A. P. Simpson, "Phonetic differences between male and female speech," in Language and Linguistics Compass, vol. 3-2, pp. 621-640, 2009.

[16] R. Vergin, A. Farhat, D. O'Shaughnessy, "Robust gender-dependent acoustic-phonetic modeling in continuous speech recognition based on a new automatic male/female classification," in Spoken Language, vol. 2, pp. 1081-1084, 1996.

[17] E. S. Parris, M. J. Carey, "Language independent gender identification," in Proceedings of ICASSP 1996, vol. 2, pp. 685-688, 1996.

[18] I. Shafran, M. Riley, and M. Mohri, "Voice signatures," in Proc. IEEE Automatic Speech Recognition and Understanding Workshop, pp. 31–36, 2003.

[19] H. Harb, and L. Chen, "Gender identification using a general audio classifier," in Proceedings of ICME 2003, pp. 733–736, 2003.

[20] Y. Hu, D. Wu, A. Nucci, "Pitch-based gender identification with two-stage classification," in Security and Communications Networks, vol. 5, pp. 211–225, 2012.

[21] M. Li, K. Han, J. Kyu, and S. Narayanan, "Automatic speaker age and gender recognition using acoustic and prosodic level information fusion," in Journal of Computer Speech &Language, vol. 27-1, pp. 151–167, 2013.

[22] S. Safavi, P. Jancovic, M. J. Russell, and M. J. Carey, "Identification of gender from children's speech by computers and humans," in Proceedings of Interspeech 2013, Lyon, France, pp. 2440–2444, 2013.

[23] K. Chachadi, S.R. Nirmala, "Gender Recognition from Speech Signal Using 1-D CNN," In: Gunjan, V.K., Zurada, J.M. (eds) Proceedings of the 2nd International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications. Lecture Notes in Networks and Systems, vol 237. Springer, Singapore, 2022.

[24] A. Krizhevsky, I. Sutskever, and G.E. Hinton, "Image Net Classification with Deep Convolutional Neural Networks," Advances in Neural Information Processing Systems, 5, 2012, 1106-1114.

[25] V. Panayotov, G. Chen, D. Povey and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015, pp. 5206-5210.

[26] I. Goodfellow, I. et al., "Generative adversarial nets," Advances in neural information processing systems, 2014, pp. 2672–2680.

[27] S. Horn, H. Bořil, "Gender classification from speech using convolutional networks augmented with synthetic spectrograms," The Journal of the Acoustical Society of America 150, Abstract A358, 2021.

[28] K. Simonyan, A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv e-prints. 2014.

[29] C. Szegedy et al., "Deep residual learning for image recognition," In CVPR, 2016.

[30] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, "Rethinking the inception architecture for computer vision," arXiv e-prints, 2015.

[31] G. Huang, Z. Liu, L. van der Maaten, K. Q. Weinberger, "Densely connected convolutional networks," arXiv preprint, 2016.

[32] C. He, J. Chao, Y. Liu, W. Zhu, W. Du, "Data Augmentation for Deep Neural Networks Model in EEG Classification Task: A Review," Frontiers in Human Neuroscience, 15, 2021.

[33] K. Yu, W. Xu, and Y. Gong, "Deep learning with kernel regularization for visual recognition," in Proceedings of the Conference on Neural Information Processing Systems. DBLP, Vancouver, BC., 2008.

[34] S. Xie, T. Yang, X. Wang, and J. Monaghan, "Hyper-class augmented and regularized deep learning for fine-grained image classification," in Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition, Boston, MA, 2015.

[35] C. Shorten and T. M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning," Journal of Big Data Vol. 6, 2019.

[36] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov Dropout, "A Simple Way to Prevent Neural Networks from Overfitting," Journal of Machine Learning Research, 15(56), 1929-1958, 2014.

[37] S. Ioffe, C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," In ICML, 2015.

[38] L. Shao, F. Zhu, X. Li, "Transfer learning for visual categorization: a survey," IEEE Trans Neural Netw Learn Syst. 2015, 26(5), 1019—34.

[39] D. Erhan, Y. Bengio, A. Courville, P.-A. Manzagol, P. Vincent, S. Bengio, "Why does unsupervised pre-training help deep learning?," J Mach Learn Res. 2010, 11:625–60.

[40] M. Palatucci, D. Pomerleau, G.E. Hinton, T. M. Mitchell, "Zero-shot learning with semantic output codes," In NIPS, 2009.

[41] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, "Gradient-based learning applied to document recognition," Proc IEEE. 1998;86(11):2278–324.

[42] A. Galdran et al., "Data-Driven Color Augmentation Techniques for Deep Skin Image Analysis," arXiv Prepr. arXiv170303702, 2017.

[43] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random Erasing Data Augmentation," arXiv Prepr. arXiv170804896, 2017.

[44] J. Jin, A. Dundar, and E. Culurciello, "Robust convolutional neural networks under adversarial noise," arXiv Prepr. arXiv151106306, 2015.

[45] A. Mikolajczyk and M. Grochowski, "Data augmentation for improving deep learning in image classification problem," 2018 International Interdisciplinary PhD Workshop (IIPhDW), 2018, pp. 117-122.

[46] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," arXiv Prepr. arXiv160505396, 2016. arXiv160505396, 2016.

[47] Ledig et al., "Photo-realistic single image super-resolution using a generative adversarial network," arXiv Prepr., 2016.

[48] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," arXiv Prepr., 2017.

[49] H. Wu, S. Zheng, J. Zhang, and K. Huang, "GP-GAN: Towards realistic high-resolution image blending," arXiv170307195, 2017.

[50] R. A. Yeh, C. Chen, T. Y. Lim, A. G. Schwing, M. Hasegawa-Johnson, and M. N. Do, "Semantic image inpainting with deep generative models," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5485–5493.

[51] X. Yi, E. Walia, P. Babyn, "Generative adversarial network in medical imaging: a review," arXiv preprint. 2018.

[52] K. Sjolander, J. Beskow, "WaveSurfer - an open source speech tool," in Proc. of ICSLP'00, volume 4, 464–467, Beijing, China, 2000.