Overcoming Data Scarcity in Speaker Identification: Dataset Augmentation with Synthetic MFCCs via Character-level RNN

Jordan J. Bird[1] Diego R. Faria[2] Cristiano Premebida[3] Anikó Ekárt[4] Pedro P. S. Ayrosa[5]

***Abstract*—Autonomous speaker identification suffers issues of data scarcity due to it being unrealistic to gather hours of speaker audio to form a dataset, which inevitably leads to class imbalance in comparison to the large data availability from non-speakers since large-scale speech datasets are available online. In this study, we explore the possibility of improving speaker recognition by augmenting the dataset with synthetic data produced by training a Character-level Recurrent Neural Network on a short clip of five spoken sentences. A deep neural network is trained on a selection of the Flickr8k dataset as well as the real and synthetic speaker data (all in the form of MFCCs) as a binary classification problem in order to discern the speaker from the Flickr speakers. Ranging from 2,500 to 10,000 synthetic data objects, the network weights are then transferred to the original dataset of only Flickr8k and the real speaker data, in order to discern whether useful rules can be learnt from the synthetic data. Results for all three subjects show that fine-tune learning from datasets augmented with synthetic speech improve the classification accuracy, F1 score, precision, and the recall when applied to the scarce real data vs non-speaker data. We conclude that even with just five spoken short sentences, data augmentation via synthetic speech data generated by a Char-RNN can improve the speaker classification process. Accuracy and related metrics are shown to improve from around 93% to 99% for three subjects classified from thousands of others when fine-tuning from exposure to 2500-1000 synthetic data points. High F1 scores, precision and recall also show that issues due to class imbalance are also solved.**

***Index Terms*—Data Augmentation, Speaker Identification, Speech Recognition, Generative Models, Human-robot Interaction, Autonomous Systems**

I. Introduction

Although there is a large amount of speech-related audio data available in the form of public datasets, autonomous speaker classification suffers issues of data scarcity due to users understandably unwilling to provide multiple hours of speech to form a dataset which inevitably leads to a heavy class imbalance. Methods such as weighting of classes during the learning process often help with the issues posed by unbalanced datasets, but this can also be detrimental depending on the severity of the underrepresented classes. This is one of the reasons that autonomous machines, for example Pepper and Nao shown in Fig. 1 often have the ability of speech recognition in the form of transcription of words and phrases but do not possess the ability to classify a specific speaker in the form of a biometric measurement.

Fig. 1: Pepper (left) and Nao (right) are state-of-the-art robots that can perform general speech recognition but not speaker classification.

A relatively new idea based on the development of generative models is that of dataset augmentation. That is, to learn rules within data in order to be able to produce new data that bares similarity. The most famous example of this, at the time of writing, is the field of *'AI Art'* where models such as the Generative Adversarial Network learn to generalise a set of artworks in order to produce new images. Though experiments like this are the most famous, dataset augmentation is a rapidly growing line of thought in multiple fields, the question asked is *"can the synthetic data produced by a generative model aid in the classification process of the original data?"*. If this is possible, then problems encountered due to class imbalance and under-representation may possibly be mitigated by exposing algorithms to synthetic data that has been produced by a generative model, based on learning from a limited set of scarce data points. In this work, in terms of contribution, we perform the first benchmarks of data augmentation fine-tune learning of Mel-Frequency Cepstral Coefficients (MFCCs) for speaker classification[1].

[1] and [2] are with ARVIS Lab, Aston University, Birmingham, UK. Emails: {birdj1, d.faria}@aston.ac.uk

[3] is with the Institute of Systems and Robotics, Department of Electrical and Computer Engineering, University of Coimbra, Coimbra, Portugal. Email: cpremebida@isr.uc.pt

[4] is with the School of Engineering and Applied Science, Aston University, Birmingham, UK. Email: a.ekart@aston.ac.uk

[5] is with the Department of Computer Science, Universidade Estadual de Londrina, Londrina, Brazil. Email: ayrosa@uel.br

[1] to the best of our knowledge and based on literature review.

These original findings and results support the hypothesis that synthetic MFCCs are useful in improving the classification of an unbalanced speaker classification dataset when knowledge gained from deep learning is transferred from the augmented data to the original set in question. This is attempted for 2,500, 5,000, 7,500 and 10,000 synthetic data objects for three subjects from the United Kingdom, Republic of Ireland, and the United States of America.

The remainder of this work is as follows. Firstly, Section II explores the background philosophy of the processes followed by this work and related experiments as well as discussing the state-of-the-art in the field. Section III outlines the Method followed including data collection, synthetic data generation, feature extraction to form datasets, and the learning processes to discern whether augmentation of MFCCs improves speaker recognition. The results of the experiments are then discussed in Section IV before future work is outlined and conclusions drawn in Section V.

II. Background and Related Work

A. Speaker Identification

Speaker Identification, also known as recognition or verification, is a pattern recognition classification task in which an individual's voice data is classified on a personal level i.e., *"is person A speaking?"* [1]. The task is useful in multiple domains, for Human-robot Interaction [2], Forensics [3] and Biometrics [4]. Identifying speakers has shown to be a relatively easy problem when a small sample size are defined, for example it is possible to perfectly classify database of 21 speakers' MFCC data extracted from audio [5]. On the other hand, researchers have pointed out an open issue in the state of the art where far more data is present, noting the then much more difficult speaker identification problem [6]–[8]. In this work, we attempt to improve the classification process of a speaker when many thousands of alternative speakers are present, through pattern matching against a large-scale dataset. The idea behind this as well as related work are further explored in Section II-B.

B. Dataset Augmentation through Synthesis

Many philosophical, psychological, and sociological studies have explored the idea of learning from imagined situations and actions [9]–[12]. Researchers argue that the ability of imagination is paramount in the learning process by improving abilities and skills through visualisation and logical dissemination. Research has also shown that imagined situations are not a perfect reflection their counterparts in reality [13]–[15]. The conclusion thus is that humans regularly learn from imagined data that does not truly reflect reality, and yet, this process is important for effective learning regardless of how realistic the imagination is, or most importantly, isn't.

The idea of data synthesis and dataset augmentation for fine-tune learning on basis data is generally inspired by the above psychological phenomenon. This is the process of generating (or *'imagining'*) new data that is inspired by the real data. Although the data is not a reflection of the basis dataset, since it does not technically exist, the idea is that patterns and rules within the basis data will be further reflected and explored in a more abstract sense in the synthetic data, and that this exercise can further improve learning processes when applied to the original basis dataset prior to any synthesis or augmentation. This idea in machine learning is a very young field of thought, becoming prominent only during the latter part of the 2010's where success has been shown in several preliminary experiments.

Though the state-of-the-art is young, there are multiple published works that show the philosophy behind this experiment in action. In Xu et al., researchers found that data augmentation leads to an overall best F-1 score for relation classification of the SemEval dataset when implemented as part of the training process for a Recurrent Neural Network [16]. Augmentation was performed by leveraging the direction of relation. A related experiment shows that NLP-based word augmentation helps to improve classification of sentences by both CNN and RNN models [17]. Of the small number of works in the field, many focus on medical data since many classification experiments suffer from an extreme lack of data. In Frid-Adar et al., researchers showed that the classification of liver lesions could be improved by also generating synthetic images of lesions using a convolutional generative adversarial network [18]. Following this, Shin et al., argued the same hypothesis for the image classification of both Alzheimer's Disease via neuroimaging and multimodal brain tumour image segmentation via a set of differing MRI images [19].

In terms of audio, related works have also argued in favour of the hypothesis being tested in this experiment. A closely related work showed that acoustic scene classification of mel-spectrograms could be improved through synthetic data augmentation [20]. Models such as Tacotron [21] learn to produce audio spectrograms from training data in order to perform realistic text-to-speech from either textual representation or internationally recognised phonemes. In terms of the problem faced by this study, speaker recognition, limited work with i-vector representations of utterances have shown promise in terms of classification after augmentation has been performed via a GAN [22].

In this work, we implement a Character-level RNN in order to generate synthetic speech data by learning from the speaker's utterances. Character-level RNNs have shown to be effective when generating natural written text [23], [24], composing music [25], and creating artwork [26]. More importantly related to this study, RNNs have also been effective in generating accurate timeseries [27] and moreover, also MFCC data [28]. Recurrence in deep learning has proved to be revolutionary in the field of speech processing [29]–[31]. Most importantly for the idea behind this work, the

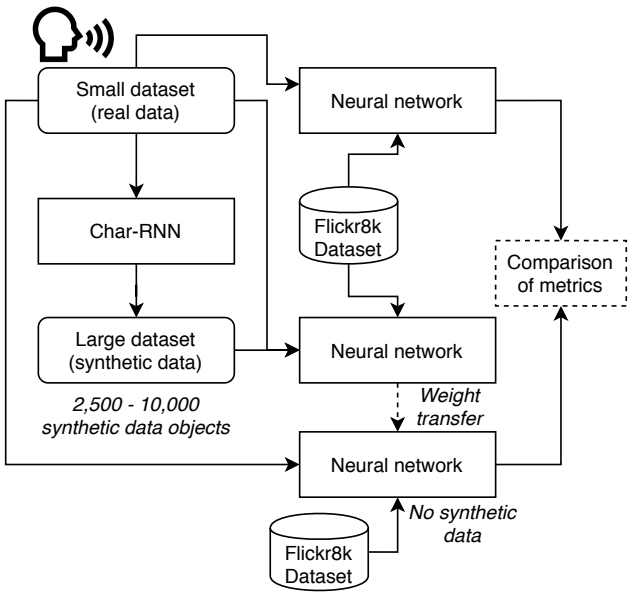


Fig. 2: Overall diagram of the experiment in which a neural network is compared to another which has been exposed to synthetic data prior to weight transfer to the clean dataset. Note that metrics regarding models trained with synthetic data are not considered.

technology has shown to be likewise as useful for the synthesis of new speech based on statistical rules within audio [28], [32], [33]. The idea is that we generate synthetic speech based on short utterances by a subject, and attempt to increase the ability of identification by also learning from the synthetic data generated by an RNN. Should this be possible, it would reduce data requirements by introducing similar speech autonomously, without the need of more extensive audio recordings.

Though some works have considered human speech in the related state-of-the-art, this work is the first preliminary exploration of synthetic data augmentation in MFCC classification for speaker identification, to the best of our knowledge.

III. Method

In this section, the method of the experiment is described. Figure 2 shows a diagram of the experimental method. To gain a set of results, three networks are trained; without synthetic data, with synthetic data, and a third without synthetic data but with the weights transferred from training when exposed to synthetic data. Thus, the comparison of the first and third models are performed since they derive directly comparable results.

A. Real and Synthetic Data Collection

The data for each experiment is split into a binary classification problem. Class *0* denotes *'not the speaker'* whereas class *1* denotes *'the speaker'*.

In order to gather a large corpus of speech for class 0, the Flickr8k dataset is gathered [34]. The dataset contains 40,000 spoken captions of 8,000 images by many speakers (unspecified by dataset authors). The process in subsection III-B is followed to generate features, and 100,000 data objects are selected at random. 50,000 of the data objects are selected in blocks of 1,000 and the remaining 50,000 are selected at random - this produces a set populated by lengthier spoken text as well as short samples of many thousands of speakers additionally.

To gather real data for class 1, three subjects as observed in Table I were asked to speak five random *Harvard Sentences*, based on the *IEEE recommended practice for speech quality measurements* [35]. This short process gathers several seconds of speech in a user-friendly manner. Users are asked to record the data via their smartphone microphone, subjects 1 and 2 used an iPhone 7 whereas subject 3 used a Samsung Galaxy S7. Although the same data is provided by all three subjects, subjects 2 and 3 spoke at a much quicker pace and thus provided far fewer data objects than subject 1.

Synthetic data for class 1 is generated by a Character level Recurrent Neural Network (Char-RNN) [36], [37], topology of the network for subject 1 is shown in Table II. The Char-RNN learns to model the probability distribution of the next character in a sequence after observing a sequence of previous characters, where previous characters are those that the RNN has also generated. By performing this one character at a time, the model initially learns the CSV formatting of the dataset (26 comma separated numerical values followed by class label *'1'* and a line break character) and then learns to form MFCC data based on observing the provided dataset.

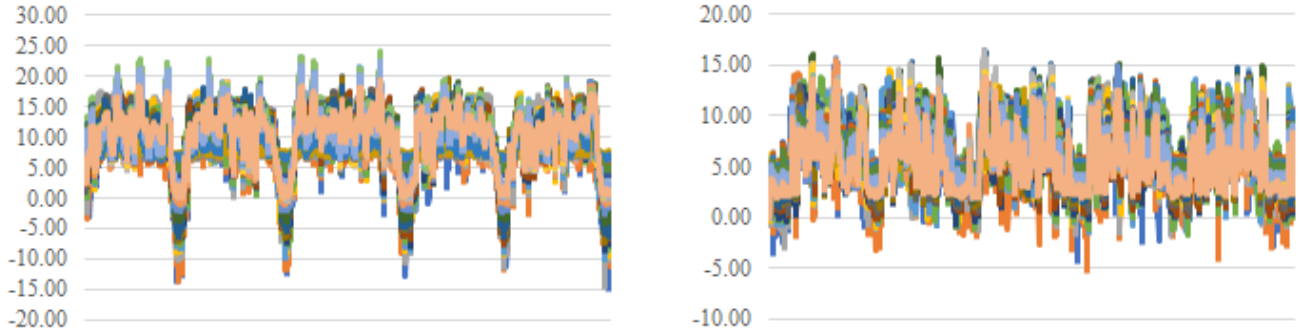
An RNN is trained for each individual subject's MFCC data (see subsection III-B) for 100 epochs before producing 10,000 synthetic data objects. This is approximately a sequence of 2,000,000 characters for each subject. An example of some data generated can be seen in Fig. 3, what seems like sound wave behaviour can be observed in the synthetic data but the nature is also noticeably different. Behaviours such as the peaks observed in the synthetic data may aid in classification of real data through augmentation, provided there are useful patterns within the probability distribution observed from the real data. This argues the need for fine-tune learning rather than transfer, since this would allow the neural network to then discard the information in the synthetic data that is unnatural, but could possibly carry forward useful rules within the nature of the generative model output.

B. Feature Extraction

The non-stationary nature of audio poses a difficult classification problem when single data points are considered [38], [39]. To overcome this, temporal statistical features are extracted from the wave. In this work, we extract the first 26 Mel-Frequency Cepstral Coefficients (MFCC) [40], [41] of the audio clips through a set of sliding windows of 0.025 seconds in length at a step of 0.01 seconds.
The MFCC extraction process is as follows:

TABLE I: Information regarding the data collection from the three subjects

Subject	Sex	Age	Nationality	Dialect/Accent	Time Taken (s)	Data Objects Captured
1	M	23	British	Birmingham	24	4978
2	M	24	American	Tampa, Florida	13	2421
3	F	28	Irish	Dublin	12	2542
Flickr8K						100,000



(a) 2500 Real MFCCs from Subject 1

(b) 2500 Synthetic MFCCs from Subject 1

Fig. 3: Two sets of values from 26 MFCCs for 2500 time windows for subject 1, one is real whereas the second is generated by the Char-RNN. A difference in patterns can be seen between the two since synthetic human speech is imperfect. X axis is temporal (each 0.025s window) and the Y axis is the MFCC value.

TABLE II: Topology of the Character-level Recurrent Neural Network

Layer	Output	Parameters
Embedding	(16, 64, 512)	7680
CuDNN LSTM	(16, 64, 256)	788,480
Dropout (0.2)	(16, 64, 256)	0
CuDNN LSTM	(16, 64, 256)	526,336
Dropout (0.2)	(16, 64, 256)	0
CuDNN LSTM	(16, 64, 256)	526,336
Dropout (0.2)	(16, 64, 256)	0
Time Distributed (size of vocabulary)	(16, 64, 15)	3855
Softmax	(16, 64, 15)	0

1) The Fourier Transform (FT) of the time window data ω is calculated:

$$X(j\omega) = \int_{-\infty}^{\infty} x(t)e^{-j\omega t}dt. \quad (1)$$

2) The powers from the FT are mapped to the Mel scale, the psychological scale of audible pitch [42] via a triangular temporal window.
3) The Mel-Frequency Cepstrum (MFC), or power spectrum of sound, is considered and logs of each of their powers are taken.
4) The derived Mel-log powers are treated as a signal, and a Discrete Cosine Transform (DCT) is measured. This is given as:

$$X_k = \sum_{n=0}^{N-1} x_n cos\left[\frac{\pi}{N}(n+\frac{1}{2})k\right] k = 0, ..., N-1, \quad (2)$$

where x is the array of length N, k is the index of the output coefficient being calculated, where N real numbers $x_0...x_{n-1}$ are transformed into the N real numbers $X_0...X_{n-1}$ by the formula.

The amplitudes of the spectrum are known as the MFCCs. The resultant data then provides a mathematical description of wave behaviour in terms of sounds, each data object made of 26 attributes produced from the sliding window are then treated as the input attributes for the neural networks.

This process is performed for all of the selected Flickr8K data as well as the real data recorded from the subjects. The MFCC data from each of the three subjects' audio recordings is used as input to the Char-RNN generative model.

C. Speaker Classification Learning Process

Datasets are organised into the following for each subject:

1) Flickr data + recorded audio
2) Flickr data + recorded audio + 2,500 synthetic data
3) Flickr data + recorded audio + 5,000 synthetic data
4) Flickr data + recorded audio + 7,500 synthetic data
5) Flickr data + recorded audio + 10,000 synthetic data

A baseline is given through the classification of set 1. Following this, models are trained on sets 2-5 in order to produce models that have been exposed to the base dataset as well as synthetic data produced by the subject's RNN model. Finally, the results are gathered by applying the model weights trained by models 2-5 and applying each of them individually to set 1 through a method of fine-tune learning. Should the classification metrics of set 1 be improved by introducing weights trained on sets 2-5 then this supports the hypothesis

TABLE III: Classification metrics for the three subjects with regards to fine-tune learning from synthetic data (scores are given for transfer learning, NOT classification of synthetic data)

Subject	Synthetic Data	Metrics			
		Accuracy	*F1*	*Precision*	*Recall*
1	***0***	93.57	0.94	0.93	0.93
	2500	98.31	0.98	0.98	0.98
	5000	98.56	**0.99**	**0.99**	0.98
	7500	**99.03**	**0.99**	**0.99**	**0.99**
	10000	98.33	0.98	0.98	0.98
2	***0***	95.13	0.95	0.95	0.95
	2500	98.43	0.98	0.98	0.98
	5000	**99.19**	**0.99**	**0.99**	**0.99**
	7500	99.11	**0.99**	**0.99**	**0.99**
	10000	97.37	0.97	0.97	0.97
3	***0***	96.58	0.97	0.97	0.97
	2500	97.77	0.97	0.97	0.97
	5000	97.83	0.98	0.98	0.98
	7500	98.35	0.98	0.98	0.98
	10000	**98.83**	**0.99**	**0.99**	**0.99**

that synthetic data allows for better classification of speaker. This process is shown as a flow diagram in Fig. 2, note that the synthetic data is not part of the two models that are compared to derive results, rather, they provide weights to be transferred to the third network. Thus, the two networks compared are trained with identical data, and differ only in terms of starting weight distribution for fine-tune learning. The hyperparameters and topology of the deep neural network are selected based on an evolutionary search approach that three deep layers of 30, 7, and 29 hidden neurons were a strong solution for the classification of MFCC attributes. Activation functions of the layers are ReLu and the ADAM optimiser [43] is used. Training is not limited to a set number of epochs, rather, early stopping is introduced at a threshold of 25 epochs with no improvement of ability before training is ceased. This therefore allows all networks to stabilise to an asymptote. Classification errors are weighted by the prominence of the class in the dataset.

All of the deep learning experiments performed in this work were executed on an Nvidia GTX980Ti GPU.

IV. Preliminary Results

The results for the three subjects can be seen in Table III. It is shown that introducing synthetic data and then transferring weights to the non-synthetic dataset network improves over no data augmentations (0 in column 2) in every case for all subjects. That is, all transfer networks outperform all non-transfer networks for each of the subjects. In each of the 12 fine-tuning experiments, classification metrics were shown to improve when the learnt knowledge was applied to the non-synthetic dataset which argues in favour of the hypothesis that the false data produced by the RNN helps to improve the speaker classification process and overcome difficulties faced by data scarcity and imbalance. Interestingly, the two male subjects hit peak performance at 5,000 to 7,500 synthetic data objects, whereas the female subject peaks at 10,000 which suggests the possibility of difference based on either gender or accent, which must be explored further in order to identify the cause (providing that it is not a fluke occurrence). This should be performed with a larger range of subjects in order to show why statistical differences may occur for improvement of classification ability. The performance for the first two subjects seemingly begins to decrease at fine-tuning from exposure to 10,000 synthetic data objects, suggesting that the generative model could be improved to prevent confusion in the model. In terms of the best improvements to classification accuracy, Subject 1 increased by 5.46% (93.57% to 99.03%) with the introduction of transfer learning from 7,500 synthetic data objects. Subject 2 increased by 4.06% (95.13% to 99.19%) by transfer learning from 5,000 synthetic data objects, and classification of Subject 3 was improved by 2.25% (96.58% to 98.83%) with transfer learning from 10,000 synthetic data objects. On average, this is a classification accuracy improvement of 3.92%. Also observed in Table III are improvements to the F_1 score, precision and recall metrics for each of the models when transfer learning from synthetic datasets.

V. Future Work and Conclusion

Since this work has provided argument in favour of the hypothesis that exposing a speech classification network to synthetic data improves speaker recognition, further work is enabled to explore this in more detail. A large limitation to the RNN was the time spent on simply learning the format of the data, that is, comma separated numerical values strictly 26 in length before being followed by a new line character. Models such as a Generative Adversarial Network (GAN) could have these rules as standard (i.e., 26 Generator outputs, 26 Discriminator inputs, 1 Discriminator output) which enables the learning to focus purely on values of attributes and their relationships with one another as well as the class label. As an extension, this experiment should be repeated with data produced by a GAN in order to compare the two methods of data synthesis.

Additionally, more subjects should be considered in future for a wider range of languages and locales for further comparison. This study was somewhat ranged with American, Irish and English subjects but further exploration should be performed in order to discern whether the effects of augmentation may change for some accents, as well as the effects that are observed should the speakers and comparison dataset speakers use a language other than English.

On a more generalised view of augmentation for learning, literature review revealed that the majority of work was performed only in the latter part of the last decade. There are many fields of machine learning in which augmentation has either been explored only slightly or even not at all, and as such cross-field co-operation is needed to further exploit the possibilities of generative augmentation processes.

To conclude, all of the experiments in this work that were augmented to any extent by synthetic data then had a measurably better classification ability of the original dataset

when compared to the learning process on said original data. This preliminary work enables much future exploration in terms of both learning models and application of findings since the issues arising from dataset imbalance are somewhat mitigated by exposure to new data.

REFERENCES

[1] A. Poddar, M. Sahidullah, and G. Saha, "Speaker verification with short utterances: a review of challenges, trends and opportunities," *IET Biometrics*, vol. 7, no. 2, pp. 91–101, 2017.

[2] E. Mumolo and M. Nolich, "Distant talker identification by nonlinear programming and beamforming in service robotics," in *IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing*, pp. 8–11, 2003.

[3] P. Rose, *Forensic speaker identification*. cRc Press, 2002.

[4] N. K. Ratha, A. Senior, and R. M. Bolle, "Automated biometrics," in *International Conference on Advances in Pattern Recognition*, pp. 447–455, Springer, 2001.

[5] M. R. Hasan, M. Jamil, M. Rahman, *et al.*, "Speaker identification using mel frequency cepstral coefficients," *variations*, vol. 1, no. 4, 2004.

[6] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017.

[7] S. Yadav and A. Rai, "Learning discriminative features for speaker identification and verification.," in *Interspeech*, pp. 2237–2241, 2018.

[8] H. Zeinali, S. Wang, A. Silnova, P. Matějka, and O. Plchot, "But system description to voxceleb speaker recognition challenge 2019," *arXiv preprint arXiv:1910.12592*, 2019.

[9] K. Egan, "Memory, imagination, and learning: Connected by the story," *Phi Delta Kappan*, vol. 70, no. 6, pp. 455–459, 1989.

[10] G. Heath, "Exploring the imagination to establish frameworks for learning," *Studies in Philosophy and Education*, vol. 27, no. 2-3, pp. 115–123, 2008.

[11] P. MacIntyre and T. Gregersen, "Emotions that facilitate language learning: The positive-broadening power of the imagination," 2012.

[12] K. Egan, *Imagination in teaching and learning: The middle school years*. University of Chicago Press, 2014.

[13] D. Beres, "Perception, imagination, and reality," *International Journal of Psycho-Analysis*, vol. 41, pp. 327–334, 1960.

[14] E. F. Loftus, "Creating false memories," *Scientific American*, vol. 277, no. 3, pp. 70–75, 1997.

[15] H. L. Roediger III, D. A. Balota, and J. M. Watson, "Spreading activation and arousal of false memories.," 2001.

[16] Y. Xu, R. Jia, L. Mou, G. Li, Y. Chen, Y. Lu, and Z. Jin, "Improved relation classification by deep recurrent neural networks with data augmentation," *arXiv preprint arXiv:1601.03651*, 2016.

[17] S. Kobayashi, "Contextual augmentation: Data augmentation by words with paradigmatic relations," *arXiv preprint arXiv:1805.06201*, 2018.

[18] M. Frid-Adar, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "Synthetic data augmentation using gan for improved liver lesion classification," in *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pp. 289–293, IEEE, 2018.

[19] H.-C. Shin, N. A. Tenenholtz, J. K. Rogers, C. G. Schwarz, M. L. Senjem, J. L. Gunter, K. P. Andriole, and M. Michalski, "Medical image synthesis for data augmentation and anonymization using generative adversarial networks," in *International workshop on simulation and synthesis in medical imaging*, pp. 1–11, Springer, 2018.

[20] J. H. Yang, N. K. Kim, and H. K. Kim, "Se-resnet with gan-based data augmentation applied to acoustic scene classification," in *DCASE 2018 workshop*, 2018.

[21] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, *et al.*, "Tacotron: Towards end-to-end speech synthesis," *arXiv preprint arXiv:1703.10135*, 2017.

[22] J.-T. Chien and K.-T. Peng, "Adversarial learning and augmentation for speaker recognition.," in *Odyssey*, pp. 342–348, 2018.

[23] D. Pawade, A. Sakhapara, M. Jain, N. Jain, and K. Gada, "Story scrambler–automatic text generation using word level rnn-lstm," *International Journal of Information Technology and Computer Science (IJITCS)*, vol. 10, no. 6, pp. 44–53, 2018.

[24] L. Sha, L. Mou, T. Liu, P. Poupart, S. Li, B. Chang, and Z. Sui, "Order-planning neural text generation from structured data," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[25] D. Eck and J. Schmidhuber, "Finding temporal structure in music: Blues improvisation with lstm recurrent networks," in *Proceedings of the 12th IEEE workshop on neural networks for signal processing*, pp. 747–756, IEEE, 2002.

[26] K. Gregor, I. Danihelka, A. Graves, D. J. Rezende, and D. Wierstra, "Draw: A recurrent neural network for image generation," *arXiv preprint arXiv:1502.04623*, 2015.

[27] T. Senjyu, A. Yona, N. Urasaki, and T. Funabashi, "Application of recurrent neural network to long-term-ahead generating power forecasting for wind power generator," in *2006 IEEE PES Power Systems Conference and Exposition*, pp. 1260–1265, IEEE, 2006.

[28] X. Wang, S. Takaki, and J. Yamagishi, "An rnn-based quantized f0 model with multi-tier feedback links for text-to-speech synthesis.," in *INTERSPEECH*, pp. 1059–1063, 2017.

[29] S. Fernández, A. Graves, and J. Schmidhuber, "An application of recurrent neural networks to discriminative keyword spotting," in *International Conference on Artificial Neural Networks*, pp. 220–229, Springer, 2007.

[30] Y. He, T. N. Sainath, R. Prabhavalkar, I. McGraw, R. Alvarez, D. Zhao, D. Rybach, A. Kannan, Y. Wu, R. Pang, *et al.*, "Streaming end-to-end speech recognition for mobile devices," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6381–6385, IEEE, 2019.

[31] H. Sak, A. W. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," 2014.

[32] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating rnn-based speech enhancement methods for noise-robust text-to-speech.," in *SSW*, pp. 146–152, 2016.

[33] H. Tachibana, K. Uenoyama, and S. Aihara, "Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4784–4788, IEEE, 2018.

[34] D. Harwath and J. Glass, "Deep multimodal semantic embeddings for speech and images," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 237–244, IEEE, 2015.

[35] E. Rothauser, "Ieee recommended practice for speech quality measurements," *IEEE Trans. on Audio and Electroacoustics*, vol. 17, pp. 225–246, 1969.

[36] I. Sutskever, J. Martens, and G. E. Hinton, "Generating text with recurrent neural networks," in *Proceedings of the 28th international conference on machine learning (ICML-11)*, pp. 1017–1024, 2011.

[37] A. Graves, "Generating sequences with recurrent neural networks," *arXiv preprint arXiv:1308.0850*, 2013.

[38] Z. Xiong, R. Radhakrishnan, A. Divakaran, and T. S. Huang, "Comparing MFCC and mpeg-7 audio features for feature extraction, maximum likelihood HMM and entropic prior HMM for sports audio classification," in *IEEE Int. Conference on Acoustics, Speech, and Signal Processing (ICASSP).*, vol. 5, 2003.

[39] J. J. Bird, E. Wanner, A. Ekárt, and D. R. Faria, "Phoneme aware speech recognition through evolutionary optimisation," in *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, pp. 362–363, 2019.

[40] L. Muda, M. Begam, and I. Elamvazuthi, "Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques," *arXiv preprint arXiv:1003.4083*, 2010.

[41] M. Sahidullah and G. Saha, "Design, analysis and experimental evaluation of block based transformation in mfcc computation for speaker recognition," *Speech Communication*, vol. 54, no. 4, pp. 543–565, 2012.

[42] S. S. Stevens, J. Volkmann, and E. B. Newman, "A scale for the measurement of the psychological magnitude pitch," *The Journal of the Acoustical Society of America*, vol. 8, no. 3, pp. 185–190, 1937.

[43] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.