Airaksinen, Manu; Juvela, Lauri; Alku, Paavo; Räsänen, Okko

Data augmentation strategies for neural network F0 estimation

Published in:
44th IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2019; Brighton; United Kingdom; 12-17 May 2019 : Proceedings

DOI:
10.1109/ICASSP.2019.8683041

Published: 01/05/2019

Document Version
Peer-reviewed accepted author manuscript, also known as Final accepted manuscript or Post-print

DATA AUGMENTATION STRATEGIES FOR NEURAL NETWORK F0 ESTIMATION

Manu Airaksinen[⋆,†], *Lauri Juvela*[†], *Paavo Alku*[†], *Okko Räsänen*[†,‡]

[⋆]Department of Clinical Neurophysiology, University of Helsinki, Finland
[†]Department of Signal Processing and Acoustics, Aalto University, Finland
[‡]Laboratory of Signal Processing, Tampere University of Technology, Finland

ABSTRACT

This study explores various speech data augmentation methods for the task of noise-robust fundamental frequency (F0) estimation with neural networks. The explored augmentation strategies are split into additive noise and channel -based augmentation and into vocoder-based augmentation methods. In vocoder-based augmentation, a glottal vocoder is used to enhance the accuracy of ground truth F0 used for training of the neural network, as well as to expand the training data diversity in terms of F0 patterns and vocal tract lengths of the talkers. Evaluations on the PTDB-TUG corpus indicate that noise and channel augmentation can be used to greatly increase the noise robustness of trained models, and that vocoder-based ground truth enhancement further increases model performance. For smaller datasets, vocoder-based diversity augmentation can also be used to increase performance. The best-performing proposed method greatly outperformed the compared F0 estimation methods in terms of noise robustness.

Index Terms— Speech analysis, F0 estimation, noise robustness, data augmentation, deep learning

1. INTRODUCTION

Fundamental frequency (F0), which is the frequency of the quasi-periodic oscillation of the vocal folds during voiced speech production, is one of the most important features of speech. Humans can modify F0 by varying the tension of the vocal folds and the subglottal pressure [1]. The main information conveyed by F0 is related to the prosody of speech, which in non-tonal languages is used to convey information beyond the literal meaning of the message, such as emphasis, emotion, and speaker characteristics. This makes F0 a widely studied feature across all of speech science, including speech synthesis, prosody, and linguistics.

F0 estimation from recorded speech (also known as *pitch estimation* or *pitch tracking*) is thus itself a fundamental problem in speech signal processing that has a wide range of applications. The traditional approach taken in F0 estimators is based on signal processing techniques that aim to capture the periodicity of the analyzed signal frame (within the range of 30–50 milliseconds). Time domain F0 estimation methods are based on, for example, the autocorrelation function (e.g., YIN [2]) or normalized cross-correlation function (e.g., RAPT [3]). Frequency domain methods utilize, for example, the energy of the linear prediction residual harmonics (e.g., SRH [4]) or instantaneous frequency (e.g., TEMPO [5]). In addition to the methods that provide raw frame-level estimates of F0, multiple methods have been developed for candidate F0 selection and/or post-processing of the raw F0 estimates for improved robustness (e.g., pYIN [6], YAAPT [7], and Nebula [8]).

In recent years with the spread of deep learning, neural estimation of F0 has also been explored. For example, CREPE [9] has produced state-of-the-art results in generic audio pitch tracking, and single sinusoid regression [10] has improved the state-of-the-art F0 estimation performance in noisy conditions. The applicability of neural networks for noise-robust F0 estimation is easy to understand: As the neural network is trained for a regression or classification task from a signal-level input to a known target F0 output, during training the input can be masked, for example, by additive noise which makes the model learn to handle noisy inputs. The noise-corruption of the model inputs during training can be seen as a form of *data augmentation*, which is a staple method in training powerful image recognition networks [11]. However, within the field of image recognition, noise augmentation is a small subset of the available methods: Most augmentation methods manipulate the image in such ways that the individual pixels are transformed without affecting the human-interpreted contents of the image. These transformations include, for example, rotation, flipping, scaling, cropping, and translation. Within the field of audio processing, similarly motivated augmentation strategies have been proposed in [12], where time stretching, pitch shifting, dynamic range compression, and background noise augmentation were used to enhance the training of a convolutional neural network (CNN) for audio scene classification. Within the field of speech technology, vocal tract length perturbation (VTLP) [13] has been successfully used to enhance automatic speech recognition (ASR) system performance [14]. Despite these early efforts, the full potential of data augmentation within the field of speech technology has not yet been explored.

In this study, we explore the applicability of *vocoder-based* modifications for speech data augmentation for neural network estimation of F0. A vocoder is a speech analysis/synthesis system, widely used in text-to-speech synthesis [15], that transforms a speech signal into a parametric representation that can be synthesized back into a speech signal. The vocoder parameters, which include time trajectories for the F0 and spectral envelope of speech, can be manipulated to modify the prosody or even the identity of the speaker, while maintaining the linguistic information and yielding perceptually acceptable quality to human listeners. Such transformations to the speech signal could be interpreted as speech processing analogues to the image-processing augmentation strategies.

The main scope of this study is to explore the use of speech data augmentation for F0 estimation given a fixed neural network model. Section 2 describes the neural network model utilized in the experiments, and Section 3 describes the proposed augmentation schemes in detail. We divide the augmentation into two main

This research was supported by Academy of Finland grants no. 312105, 312490, 314573, and 314602.

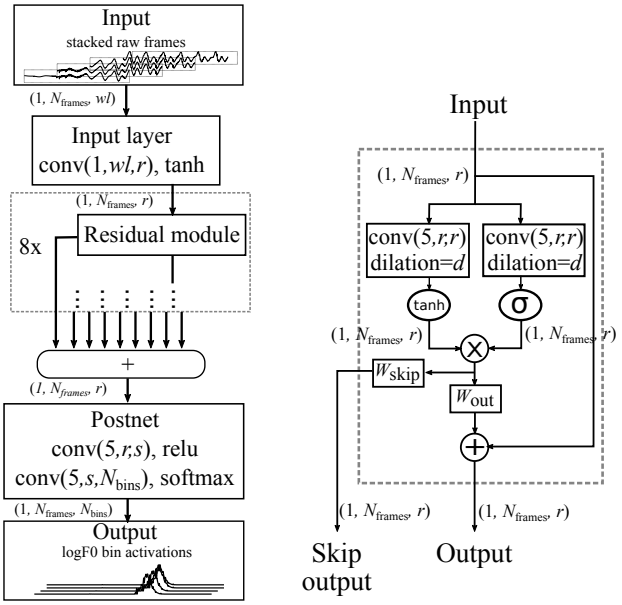
categories: *additive noise and varying channel augmentation* (Section 3.1) and *vocoder-based augmentation* (Section 3.2). The experiments, trained models, and used data sets are described in Section 4, and finally the results and discussion are presented in Sections 5 and 6, respectively.

2. NEURAL NETWORK F0 ESTIMATION METHOD

The neural network model utilized in this study is a CNN based on the WaveNet [16] architecture presented in Figure 1. Key points within the WaveNet architecture are the residual skip connections that allow error gradients to flow more efficiently during the training of deep networks, as well as the dilated convolutions that increase the receptive field of the network. The network takes in stacked raw microphone signal waveform frames as input with tensor dimensions $(1, N_{\text{frames}}, wl)$, where N_{frames} is the number of frames in the utterance to be processed and wl is the frame length ($wl = 512$ samples at F_s = 16 kHz is used within the present study). Framing is performed with a rectangular window and a 10-ms frame skip. The target outputs are one-hot vectors of evenly distributed logF0 bins between the range $[\log \text{F0}_{min}, \log \text{F0}_{max}]$ plus one bin for the voiced/unvoiced decision. In this study, we used the values $\text{F0}_{min} = 50$ Hz, $\text{F0}_{max} = 500$ Hz, and $N_{bins} = 351$.

The input layer performs a linear transformation followed by $\tanh$ activation to the input frames, yielding $r = 128$ feature channels to be processed by the residual module stack. The residual module stack consists of eight residual modules illustrated in Figure 1 (b). Each residual module consists of a gated dilated convolution operation and a skip connection that connects the module input directly into the main output of the module. The 1D convolution operation within the module is performed with varying levels of *dilations* over the frame time axis to model the time structure of F0 contours. In this study, the used dilations for the 8-layer residual module stack are $d = [1, 2, 4, 8, 1, 2, 4, 8]$, yielding (including the postnet) a receptive field of 71 frames with the selected filter length of 5 (current frame plus 35 past frames and 35 future frames) that condition each F0 estimate. The output of the convolution is passed through $\tanh$ activation and multiplied by a gating activation produced by a similar operation with the logistic sigmoid activation function. Finally, the output of the gated convolution is split into skip output and main output paths that each apply a linear transformation. The skip output is fed directly to the end of the residual module stack, whereas the main output is added back into the module skip connection to obtain the final module output.

At the end of the residual module stack, each skip connection and the final output connection (all of tensor shape $(1, N_{\text{frames}}, r)$) are combined with the addition operation (Figure 1 (a)), and a two-layer postnet with a number of channels $s = 256$ is used to predict the final logF0 bin activations that are normalized by the softmax function. The corresponding F0 estimate of the frame is given by simply taking the argmax of the activations. The total number of trainable parameters within the neural network model is 2,256,351.



(a) Overall structure (b) Residual module

Fig. 1. Block diagram of the utilized neural network model. The notation conv(x, y, z) denotes the 1D convolution operation with 'same' padding performed over the second axis of the input tensor, where x is the filter size, y the number of input channels and z the number of output channels.

3. SPEECH DATA AUGMENTATION METHODS

3.1. Noise and channel augmentation

Humans can comprehend speech under adverse conditions, where the signal-to-noise ratio (SNR) can be as low as -3 dB [17]. Furthermore, environmental noise that is common in speech communication situations is not only additive noise from another point source (e.g., a car, another speaker), but also convolutional *channel noise* (e.g., reverberation caused by the acoustic space, orientation of head w.r.t. source). To mimic these conditions that the human auditory system is adapted to, speech data is augmented both with additive noise and with convolutional channel noise. This data augmentation is conducted with random sampling from the NOISEX-92 database [18], which includes various types of generated noise signals (e.g., white, pink), as well as recorded noise samples (e.g., car, factory, babble). The channel noise augmentation is performed by convolving the speech waveform with a randomly generated impulse response.

During training of the neural network model, noise augmentation is applied once per epoch for each input utterance as follows: First, a coin is flipped separately whether to perform additive noise and/or channel augmentation. Additive noise augmentation is performed by sampling target global SNR (in dB) from uniform distribution between [-5, 15], and adding a random sample from the NOISEX database (with random starting location) to the input waveform, scaled to match the sampled SNR. Channel noise augmentation is performed by generating a random FIR filter of length 17 (F_s = 16 kHz) by sampling from a Gaussian distribution. Generated impulse response is scaled with a random gain sampled from uniform distribution between [0, 1]. Value of middle sample is set to 1.0, and the input waveform is convolved with the generated filter.

3.2. Vocoder-based augmentation

For vocoder-based augmentation, we utilized GlottDNN [19], which is a glottal vocoder based on the source-filter model of speech production [1]. GlottDNN was selected as the vocoder for the current study because it generates more realistic mixed phase characteristics to the synthesized speech, as opposed to the minimum phase waveform generation used by conventional vocoders [19]. The *source*

parameters of GlottDNN contain information about the glottal excitation (F0, harmonic-to-noise ratio, spectral tilt) and the *filter parameters* represent the vocal tract transfer function. Using these parameters, the vocoder builds the speech signal in the synthesis part frame by frame. In other words, each frame of speech is synthesized to have the target F0. This property is effective in obtaining accurate ground truth labels for F0 estimation, which we call *ground truth enhancement*. Furthermore, as the GlottDNN vocoder's parameters are easily transformable, they can be modified to change the prosody and the speaker identity. Adding new data based on this approach is called *diversity augmentation* in this study.

Ground truth enhancement

In neural estimation of F0, the ground truth F0 is needed in the training phase of the estimator. One option is to use an existing F0 estimation method for this task, in which case the F0 targets utilized in training are affected by errors made by the selected method. If these errors are small, this approach may still produce results better than the original estimator.

Another strategy is to utilize parallel recordings of speech and electroglottography (EGG) [10]. This approach, however, is limited by difficulties in recording large amounts of such parallel data, and by the fact that EGG might not give accurate F0 estimates due to, for example, imperfect attachment of the EGG electrodes or inaccuracies of automated estimation algorithms [3]. Yet another approach for obtaining ground truth labels is to use synthetic data, as is done in the CREPE method [9], where synthesized audio pitch tracks are used to train the neural model, thus circumventing the problem of erroneous labels.

The vocoder-based analysis-synthesis procedure proposed in the current study combines the benefits of using both synthetic and real data: An existing F0 estimation method is first used to extract F0 trajectories of natural speech together with the glottal vocoder-based parameterization of the signal. This is followed by vocoder-based synthesis of the same speech sample by using the initial F0 estimate. As a result, the synthesized speech signal still exhibits the natural speech parameter trajectories obtained with the vocoder, but the process enforces a consistent ground truth F0 due to the deterministically synthesized waveform. Within our experiments, the ground truth enhancement (GTE) is performed with the default settings of the GlottDNN vocoder using REAPER [20] as the F0 estimation method applied to original clean training signals.

Diversity augmentation

Gender is one of the main markers of speaker identity. Compared to female speech, male speech is characterized by relatively low F0 values (due to larger and heavier vocal folds) and lower formant frequencies of the vocal tract transfer function (due to a longer vocal tract) [1]. The vocoder parameters of a given speech sample can be transformed to change the gender identity of the sample (or perform something in between). In GlottDNN, these parameters are the F0 vector and the vocal tract warping coefficient, which are the manipulated parameters in the present study.

Our method to perform diversity augmentation (DA) is fairly simple: First, the F0 vector is modified by randomly sampling a target mean F0 for the utterance from a uniform distribution in range [100 Hz, 350 Hz]. Next, we compute the ratio ϕ between the original mean F0 and the new target, and scale the original F0 vector according to ϕ. The vocal tract length manipulation is performed by having a mismatch in the coefficient λ of the vocal tract filter between analysis and synthesis: during analysis, the coefficient is kept at $\lambda_{an} = 0.0$, and during synthesis, its value is uniformly sampled from $\lambda_{syn} \in [-0.05 - \ln\phi, 0.05 - \ln\phi]$. The effect of ϕ to the value of roughly mimics the properties of human physiology: If F0 is increased, the effective vocal tract length is, on average, proportionally decreased (i.e., the voice is made more "female"), and vice versa.

The tested DA method was kept simple in order to start from simple and easily understandable modifications. Additional DA methods could include more complex time trajectory modulations to the vocoder parameters, and/or modifying the speech rate, sharpness of the formant peaks, spectral tilt of the glottal excitation, etc.

4. EXPERIMENTS

4.1. Databases for training and evaluation

To keep the training and evaluation of the neural network models fully separate, we utilized two distinct speech corpora for training and evaluation. As a pre-processing step, all of the utilized speech corpora were resampled to a 16 kHz sampling frequency.

For training the neural network models, we utilized the full CSTR Voice Cloning Toolkit (VCTK) Corpus [21]. The CSTR VCTK corpus contains high-quality speech recordings uttered by 109 English speakers (61 female, 47 male) with various accents. Each speaker reads approximately 400 sentences from various sources, yielding a total number of 44,257 utterances with a total length of 44h2m46s. This translates to approximately 15.9 million frames of training data at a frame rate of 100 frames per second, of which approximately 36% are voiced. The size and high recording quality of the CSTR VCTK corpus makes it ideal for F0 estimator training, as conventional pitch extractors are expected to perform well with the data to obtain high-quality ground truth labels.

For evaluation, we utilized the entire Pitch Tracking Database from Graz University of Technology (PTDB-TUG) [22]. The PTDB-TUG corpus contains parallel speech and EGG recordings from 20 native English speakers (10 female, 10 male), who each read a subset of 236 sentences from the TIMIT corpus [23]. This yields a total size of 4720 recorded sentences, with a total length of 9h36min13s, of which approximately 24% are voiced. The ground-truth F0 contours provided with the database have been extracted from EGG with the RAPT method [3]. An important point in selecting the evaluation database is that it does not contain the same sentences as the training material, which means that memorizing utterance-specific dynamics is not a viable strategy for the neural network models.

4.2. Trained neural network models

For our experiments, we trained a total number of six neural F0 estimators with the architecture described in Section 2 with varying training data augmentation methods. The ground truth F0 tracks were estimated with the REAPER algorithm. For all trained methods, we randomly selected 1,000 utterances from the CSTR VCTK corpus for a validation set, and used the rest of the corpus (or its subset) for training. A full list of the trained models is presented in Table 1.

For each model, the training was performed using stochastic gradient descent (SGD) with the categorical cross-entropy loss function. We used the Adam optimizer [24] with learning rate 10^{-4}, and parameters $\beta_1 = 0.9$, $\beta_2 = 0.99$, and $\epsilon = 10^{-8}$ to perform SGD updates. We used a variable batch size where each batch contained

Table 1. Trained neural network models. NA=additive noise and varying channel data augmentation, GTE=ground truth enhancement, DA=diversity augmentation.

Model name	Augmentation method(s)	# Utterances of train data
BL0	none	43,257
BL1	NA	43,257
GteAug	NA + GTE	43,257(GTE)
DivAug	NA + GTE + DA	43,257(GTE) + 389,313(DA)
BL1 partial	NA	4,326
DivAug partial	NA + GTE + DA	4,326(GTE) + 38,934(DA)

all frames of a single utterance. Dropout with $p = 0.3$ was applied to the second layer inputs. Each method was trained for 100 epochs, and model from the epoch with the smallest validation loss was selected for the experiments.

4.3. Evaluated methods

Four existing F0 estimation techniques were selected as reference: YAAPT [7], REAPER [20], SRH [4], and CREPE [9]. YAAPT is a widely used hybrid time- and frequency domain method that can be run with varying levels of complexity. For our tests we have used the 'full' and 'very fast' versions of YAAPT. REAPER is a state-of-the-art time-domain approach for pitch tracking and glottal closure instant detection, aimed mainly for high-quality data. SRH is a frequency-domain pitch estimation method that aims for noise robustness. CREPE is a well-known neural pitch estimator, whose main application area is generic audio (e.g., music). Note that unlike the other compared methods, CREPE does not produce voicing decisions and therefore this aspect cannot be evaluated for the method.

4.4. Test setup

Methods were evaluated with corrupted speech of two noise types (white, babble) with four different global SNR levels (15, 5, 0, and -5 dB), and with clean speech. We used three standard performance metrics taken from [4]: 1) voicing decision error VDE (%), the relative number of erroneous voicing detections, 2) gross prediction error GPE (%), the relative number of gross errors ($> 20\%$) in F0 estimates within the correctly detected voiced frames, and 3) fine prediction error FPE (%), the standard deviation of non-gross errors on voiced frames (in % relative to absolute F0). For the sake of conciseness, we report average performance numbers across white and babble noise, as qualitative differences between the two were minor.

5. RESULTS

The obtained results are presented in Figure 2. For clarity, we have divided the result plots into two groups: the left column contains the results for the trained neural network models, and the right column contains the results for the evaluated methods in addition to the best performing proposed model, GteAug. By analyzing the results of the left column, we can see that any type of augmentation greatly increases VDE performance over the BL0 system. The effect of GTE further increases performance especially in the VDE and FPE categories over the otherwise second best performing system, BL1. The effect of DA is perhaps most interesting: Expanding the full CSTR VCTK corpus with DA does not increase F0 estimation performance (see DivAug vs BL1), but with limited data, the DA strategy can be seen to be beneficial (see DivAug partial vs BL1 partial).

Looking at the obtained results for the other methods, we can see that the proposed method outperforms the other systems espe-

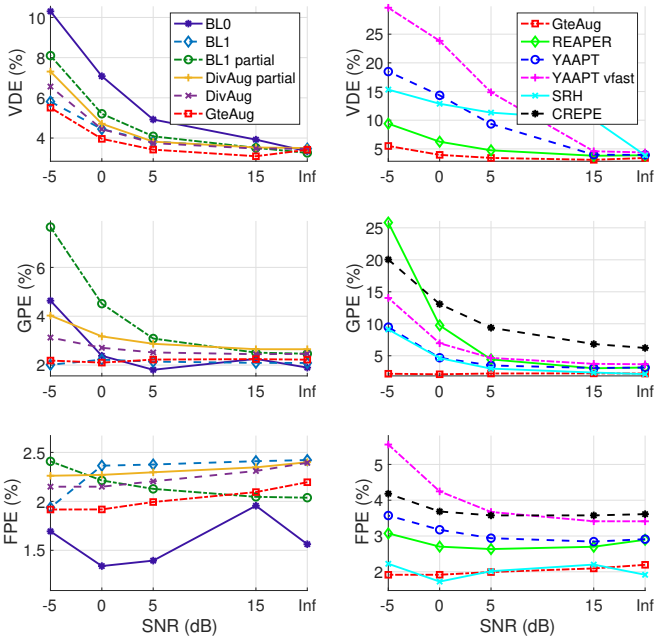


Fig. 2. Averaged test results over white and babble noise on the PTDB-TUG database.

cially in the VDE and GPE categories. Strikingly, the method seems to have a flat GPE performance even at -5 dB. The counter-intuitive behavior of the FPE curve for GteAug and SRH can be explained by the increasing number of voicing errors: as more low-energy frames, from which F0 is generally harder to detect, are classified as unvoiced when SNR decreases, the number of frames from which FPE is computed decreases. These results also compare favorably against the DNN-based results recently reported on a subset of the same PTDB-TUG corpus [10]. However, direct comparison is difficult, as the authors of [10] performed cropping of silence regions in the signals before SNR calculation. Still, GPE in [10] is always substantially larger than in GteAug across the entire SNR range, while FPE is similar to the now-proposed approach.

6. DISCUSSION

This study presented a neural network-based F0 estimation model that was trained with varying strategies of speech data augmentation. The data augmentation methods, divided into noise-based augmentation and vocoder-based augmentation categories, were presented in detail, and a total number of six different neural F0 estimation methods were trained with varying modes of data augmentation. The obtained objective results show that speech data augmentation can be beneficial in training powerful, noise robust neural network models: Vocoder-based speech analysis/synthesis can be used to enhance the ground-truth labels used in supervised training, and for limited size datasets vocoder-based diversity augmentation can be used to gain performance. This suggests that vocoder-based augmentation could be also applicable to other problems were data sparsity is more of an issue than in standard F0 estimation. Finally, compared to the pre-existing F0 estimation methods, the best performing proposed method using noise- and ground truth enhancement yields considerable performance gains under adverse noise conditions. Our future work consists of applying the proposed framework to coded telephone speech.

7. REFERENCES

[1] G. Fant, *Acoustic Theory of Speech Production*, De Gruyter, 1970.

[2] A. de Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002.

[3] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Palatal, Eds., pp. 497–518. Elsevier Science B.V., 1995.

[4] T. Drugman and A. Alwan, "Joint robust voicing detection and pitch estimation based on residual harmonics," in *Proc. Interspeech*. ISCA, 2011, pp. 1973–1976.

[5] H. Kawahara, A. de Cheveigné, and R. D. Patterson, "An instantaneous-frequency-based pitch extraction method for high-quality speech transformation: revised TEMPO in the STRAIGHT-suite," in *Proc. ICLSP*, 1998.

[6] M. Mauch and S. Dixon, "PYIN: A fundamental frequency estimator using probabilistic threshold distributions," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 659–663.

[7] S. A. Zahorian and H. Hu, "A spectral/temporal method for robust fundamental frequency tracking," *The Journal of the Acoustical Society of America*, vol. 123, no. 6, pp. 4559–4571, 2008.

[8] K. Hua, "Nebula: F0 estimation and voicing detection by modeling the statistical properties of feature extractors," in *Proc. Interspeech*, 2018, pp. 337–341.

[9] J. W. Kim, J. Salamon, P. Li, and J. P. Bello, "Crepe: A convolutional representation for pitch estimation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 161–165.

[10] A. Kato and T. Kinnunen, "Waveform to single sinusoid regression to estimate the F0 contour from noisy speech using recurrent deep neural networks," in *Proc. Interspeech*, 2018.

[11] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," *arXiv:1608.06993 [cs.CV]*, 2016.

[12] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Processing Letters*, vol. 24, no. 3, pp. 279–283, 2017.

[13] N. Jaitly and G. E. Hinton, "Vocal tract length perturbation (VTLP) improves speech recognition," in *Proc. ICML*, 2013.

[14] I. Rebai, Y. BenAyed, W. Mahdi, and J.-P. Lorré, "Improving speech recognition using data augmentation and acoustic model fusion," *Procedia Computer Science*, vol. 112, pp. 316–322, 2017.

[15] H. Zen, K. Tokuda, and A. W. Black, "Review: Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[16] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *CoRR*, vol. abs/1609.03499, 2016.

[17] J. B. Nielsen, *Assessment of speech intelligibility in background noise and reverberation*, Ph.D. thesis, Technical University of Denmark, 2009.

[18] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247–251, 1993.

[19] M. Airaksinen, L. Juvela, B. Bollepalli, J. Yamagishi, and P. Alku, "A comparison between STRAIGHT, glottal, and sinusoidal vocoding in statistical parametric speech synthesis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 9, pp. 1658–1670, 2018.

[20] D. Talkin, "REAPER: Robust epoch and pitch estimator," https://github.com/google/REAPER, 2014, Accessed: 2018-10-26.

[21] C. Veaux, J. Yamagishi, and K. MacDonald, "CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit," https://datashare.is.ed.ac.uk/handle/10283/2651, 2017, Accessed: 2018-10-26.

[22] G. Pirker, M. Wohlmayr, S. Petrik, and F. Pernkopf, "A pitch tracking corpus with evaluation on multipitch tracking scenario," in *Proc. Interspeech*, 2011, pp. 1509–1512.

[23] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "TIMIT acoustic-phonetic continuous speech corpus," in *LDC93S1 [Web Download]. Linguistic Data Consortium*, 1993.

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