SPEECH DENOISING BY PARAMETRIC RESYNTHESIS

Soumi Maiti and Michael I Mandel

The CUNY Graduate Center and Brooklyn College
New York, NY

ABSTRACT

This work proposes the use of clean speech vocoder parameters as the target for a neural network performing speech enhancement. These parameters have been designed for text-to-speech synthesis so that they both produce high-quality resyntheses and also are straightforward to model with neural networks, but have not been utilized in speech enhancement until now. In comparison to a matched text-to-speech system that is given the ground truth transcripts of the noisy speech, our model is able to produce more natural speech because it has access to the true prosody in the noisy speech. In comparison to two denoising systems, the oracle Wiener mask and a DNN-based mask predictor, our model equals the oracle Wiener mask in subjective quality and intelligibility and surpasses the realistic system. A vocoder-based upper bound shows that there is still room for improvement with this approach beyond the oracle Wiener mask. We test speaker-dependence with two speakers and show that a single model can be used for multiple speakers.

Index Terms— Speech enhancement, synthesis, vocoder

1. INTRODUCTION

The general approach of speech enhancement has been to modify a noisy signal to make it more like the clean signal [1]. The main problems for such systems are the over-suppression of the speech and under-suppression of the noise. Ideally, speech enhancement systems should remove the noise completely without decreasing the speech quality. There are, however, statistical text-to-speech (TTS) synthesis systems that can produce high-quality speech from textual inputs (e.g., [2]) by training an acoustic model to map text to the time-varying acoustic parameters of a vocoder, which then generates the speech. The most difficult part of this task, however, is predicting realistic prosody (timing information and pitch and loudness contours) from pure text.

In this paper, we propose combining these two approaches to capitalize on the strengths of each by predicting the acoustic parameters of clean speech from a noisy observation and then using a vocoder to synthesize the speech. We show that this combined system can produce high-quality and noise-free speech utilizing the true prosody observed in the noisy speech. We demonstrate that the noisy speech signal has more information about the clean speech than its transcript does. Specifically, it is easier to predict realistic prosody from the noisy speech than from text. Thus, we train a neural network to learn the mapping from noisy speech features to the acoustic parameters of the corresponding clean speech. From the predicted acoustic features, we generate clean speech using a speech synthesis vocoder. Since we are creating a clean resynthesis of the noisy signal, the output speech quality will be higher than standard speech denoising systems and completely noise-free. We refer to the proposed model as *parametric resynthesis*.

In this paper, we show that parametric resynthesis outperforms statistical TTS in terms of traditional speech synthesis objective metrics. Next we subjectively evaluate the intelligibility and quality of the resynthesized speech and compare it with a mask predicted by a DNN-based system [3] and the oracle Wiener mask [4]. We show that the resynthesized speech is noise-free and has overall quality and intelligibility equivalent to the oracle Wiener mask and exceeding that of the DNN-predicted mask. We also show that a single parametric resynthesis model can be used for multiple speakers.

2. RELATED WORK

Traditional speech synthesis systems are of two types, concatenative and parametric. In our previous works, [5, 6, 7, 8] we proposed concatenative synthesis systems for denoising speech. Though these models can generate high quality speech, they are speaker-dependent and generally require a large dictionary of speech examples from that speaker. Alternatively, the current paper utilizes a parametric speech synthesis model, which more easily generalizes to combinations of conditions not seen explicitly in training examples.

In terms of parametric resynthesis, Rethage et al. [9] built an end-to-end model to map noisy audio to explicit models of both clean speech and noise using a WaveNet-like [10] architecture. Compared to this model, our denoising system is much simpler, as it does not require an explicit model of the observed noise in order to converge and needs much less data and time to train. This simplicity comes from using the non-neural WORLD vocoder [11].

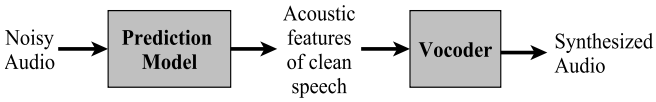

Fig. 1. Vocoder denoising model

3. MODEL OVERVIEW

Parametric resynthesis consists of two stages: prediction and synthesis as shown in Figure 1. The first stage is to train a prediction model with noisy audio features as input and clean acoustic features as output labels. The second stage is to resynthesize audio using the vocoder from the predicted acoustic features.

We use the WORLD vocoder [11] to transform between acoustic parameters and clean speech waveform. This vocoder allows both the encoding of speech audio into acoustic parameters and the decoding of acoustic parameters back into audio with very little loss of speech quality. The acoustic parameters are much easier to predict using neural network prediction models than the raw audio. We use the encoding of clean speech to generate our training targets and the decoding of predictions to generate output audio. The WORLD vocoder is incorporated into the Merlin neural network-based speech synthesis system [2], and we utilize Merlin's training targets and losses for our model.

The prediction model is a neural network that takes as input the log mel spectra of the noisy audio and predicts clean speech acoustic features at a fixed frame rate. The WORLD encoder outputs four acoustic parameters: i) spectral envelope, ii) log fundamental frequency (F0), iii) a voiced/unvoiced decision and iv) aperiodic energy of the spectral envelope. All the features are concatenated with their first and second derivatives and used as the targets of the prediction model. There are 60 features for spectral envelope, 5 for band aperiodicity, 1 for F0 and a boolean flag for the voiced/unvoiced decision. The prediction model is then trained to minimize the mean squared error loss between prediction and ground truth. This architecture is similar to the acoustic modelling of statistical TTS. We first use a feed-forward DNN as the core of the prediction model, then we use LSTMs [12] for better incorporation of context. For the feed-forward DNN, we include an explicit context of ± 4 neighboring frames.

4. EXPERIMENTS

4.1. Dataset

The noisy dataset is generated by adding environmental noise to the CMU arctic speech dataset [13]. The arctic dataset contains the same 1132 sentences spoken by four different speakers. The speech is recorded in studio environment. The sentences are taken from different texts from Project Gutenberg and are phonetically balanced. We add environmental noise from the CHiME-3 challenge dataset [14]. The noise was recorded in four different environments: street, pedestrian walkway, cafe, and bus interior. Six channels are available for each noisy file, we treat each channel as a separate noise recording. We mix clean speech with a randomly chosen noise file starting from a random offset with a constant gain of 0.95. The signal-to-noise ratio (SNR) of the noisy files ranges from -6 dB to 21 dB, with average being 6 dB. The sentences are 2 to 13 words long, with a mean length of 9 words. We mainly use a female speech corpus ("slt") for our experiments. A male ("bdl") voice is used to test the speaker-dependence of the system. The dataset is partitioned into 1000-66-66 as train-dev-test. Features are extracted with a window size of 64 ms at a 5 ms hop size.

4.2. Evaluation

We evaluate two aspects of the parametric resynthesis system. Firstly, we compare speech synthesis objective metrics like spectral distortion and errors in F0 prediction with a TTS system. This quantifies the performance of our model in transferring prosody from noisy to clean speech. Secondly, we compare the intelligibility and quality of the speech generated by parametric resynthesis (PR) against two speech enhancement systems, a DNN-predicted ratio mask (DNN-IRM) [3] and the oracle Wiener mask (OWM) [4]. The ideal ratio mask DNN is trained with the same data as PR. The OWM uses knowledge of the true speech to compute the Wiener mask and serves as an upper bound on the performance achievable by mask based enhancement systems[1].

A limitation of the proposed method is that the vocoder is not able to perfectly reproduce clean speech, so we encode and decode clean speech with it in order to estimate the loss in intelligibility and quality attributable to the vocoder alone, which we show is small. We call this system vocoder-encoded-decoded (VED). Moreover, we also measure the performance of a DNN that predicts vocoder parameters directly from clean speech as a more realistic upper bound on our speech denoising system. This is the PR model with clean speech as input, referred to as PR-clean.

4.3. TTS objective measures

First, we evaluate the TTS objective measures for PR, PR-clean, and the TTS system. We train the feedforward DNN with 4 layers of 512 neurons each with tanh activation function and the LSTM with 2 layers of width 512 each. We use adam optimization [15] and early stopping regularization. For TTS system inputs, we use the ground truth transcript of the noisy speech. As both TTS and PR are predicting acoustic features, we measure errors in the prediction via mel cepstral distortion (MCD), band aperiodicity distortion (BAPD), F0 root mean

[1] All files are available at `http://mr-pc.org/work/icassp19/`

System	Spectral Distortion		F0 measures		
	MCD (dB↓)	BAPD (dB↓)	RMSE (Hz↓)	CORR (↑)	VUV (↓)
PR-clean	2.68	0.16	4.95	0.96	2.78%
TTS (DNN)	5.28	0.25	13.06	0.71	6.66%
TTS (LSTM)	5.05	0.24	12.60	0.73	5.60%
PR (DNN)	5.07	0.19	8.83	0.93	6.48%
PR (LSTM)	**4.81**	**0.19**	**5.62**	**0.95**	**5.27%**

Table 1. TTS objective measures for single-speaker experiment: mean cepstral distortion (MCD), band aperiodicity (BAPD), root mean square error (RMSE), voiced-unvoiced error rate (VUV), and correlation (CORR). For MCD, BAPD, RMSE, and VUV lower is better (↓), for CORR higher is better (↑).

Model	Speakers		Spectral Distortion		F0 measures		
	Train	Test	MCD (dB↓)	BAPD(dB↓)	RMSE(Hz↓)	CORR(↑)	UUV(↓)
PR	slt	slt	4.81	0.19	5.62	0.95	5.27%
PR	slt+bdl	slt	4.91	0.20	8.36	0.92	6.50%
PR	bdl	bdl	5.40	0.21	9.67	0.82	12.34%
PR	slt+bdl	bdl	5.19	0.21	10.41	0.82	12.17%

Table 2. TTS objective measures for multi-speaker parametric resynthesis models compared to single speaker model.

square error (RMSE), Pearson correlation (CORR) of F0, and classification error in voiced-unvoiced decisions (VUV). The results are reported in Table 1.

Results from PR-clean show that acoustic parameters that generate speech with very low spectral distortion and F0 error can be predicted from clean speech. More importantly, we see from Table 1 that PR performs considerably better than the TTS system. It is also interesting to note that the F0 measures, RMSE and Pearson correlation are significantly better in the parametric resynthesis system than TTS. This demonstrates that it is easier to predict acoustic features, including prosody, from noisy speech than from text. We observe that the LSTM performs best and it is used in our subsequent experiments.

Evaluating multiple speaker model Next we train a PR model with speech from two speakers and test its effectiveness on each speaker's dataset. We first train two single-speaker PR models using the slt (female) and bdl (male) data in the CMU arctic dataset. Then we train a new PR model with speech from both speakers. We measure the objective metrics on both datasets to understand how well a single model can model both speakers. These objective metrics are reported in Table 2, from which we observe that the single-speaker models slightly out-perform the multi-speaker models. On the bdl dataset, however, the multi-speaker model performs better than the single-speaker model in predicting voicing decisions and in MCD. It scores the same in BAPD and F0 correlation, but does worse on F0 RMSE. These results show that the same model can be used for multiple speakers. In future work we will investigate the degree to which a single model can generalize to completely unseen speakers.

4.4. Speech enhancement objective measures

We measure objective intelligibility with short-time-objective-intelligibility (STOI) [16] and objective quality with perceptual evaluation of speech quality (PESQ) [17]. We compare the clean, noisy, VED, TTS, PR-clean speech for reference. The results are reported in Table 3.

Of the vocoder-based systems, VED shows very high objective quality and intelligibility. This demonstrates that the vocoder is able to produce high fidelity speech when it is fed with acoustic parameters that are exactly correct. The PR-clean system shows slightly lower intelligibility and quality than VED. The TTS system shows very low quality and intelligibility, but this can be explained by the fact that the objective measures compare the output to the original clean signal.

For the speech denoising systems, the oracle Wiener mask performs best, because it has access to the clean speech. While it is an upper bound on mask-based speech enhancement, it does degrade the quality of the speech from the clean by attenuating regions where there is speech present, but the noise is louder. Parametric resynthesis outperforms the predicted IRM in objective quality and intelligibility.

4.5. Subjective Intelligibility and Quality

Finally we evaluate the subjective intelligibility and quality of PR compared with OWM, DNN-IRM, PR-clean, and the ground truth clean and noisy speech. From 66 test sentences, we chose 12, with 4 sentences from each of three groups: SNR < 0 dB, 0 dB $\leq$ SNR < 5 dB, and 5 dB $\leq$ SNR. Preliminary listening tests showed that the PR-clean files sounded quite

Model	PESQ	STOI
Clean	4.50	1.00
VED	3.39	0.93
OWM	3.31	0.96
PR-clean	2.98	0.92
PR	2.43	0.87
DNN-IRM	2.26	0.80
Noisy	1.88	0.88
TTS	1.33	0.08

Table 3. Speech enhancement objective metrics: quality (PESQ) and intelligibility (STOI), higher is better for both. Systems in the top section use oracle information about the clean speech. All systems sorted by PESQ.

similar to the VED files, so we included only PR-clean. This resulted in a total of 84 files (7 versions of 12 sentences).

For the subjective intelligibility test, subjects were presented with all 84 sentences in a random order and were asked to transcribe the words that they heard in each one. Four subjects listened to the files. A list of all of the words was given to the subjects in alphabetical order, but they were asked to write what they heard. Figure 2 shows the percentage of words correctly identified averaged over all files. Intelligibility is very high ($> 90\%$) in all systems, including noisy. PR-clean achieves intelligibility as high as clean speech. OWM, PR, and noisy speech had equivalent intelligibility, slightly below that of clean speech. This shows that PR achieves intelligibility as high as the oracle Wiener mask.

The speech quality test follows the Multiple Stimuli with Hidden Reference and Anchor (MUSHRA) paradigm [18]. Subjects were presented with all seven of the versions of a given sentence together in a random order without identifiers, along with reference clean and noisy versions. The subjects rated the speech quality, noise reduction quality, and overall quality of each version between 1 and 100, with higher scores denoting better quality. Three subjects participated and results are shown in Figure 3.

From the results, we see that the PR system achieves higher noise suppression quality than the OWM, demonstrating that the output is noise-free. PR also achieves comparable overall quality to OWM and PR-clean, indicating that its performance is close to the ceiling imposed by the vocoder. This ceiling is demonstrated by the difference between PR-clean and the original clean speech. Note also that the large objective differences between PR and OWM are not present in the subjective results, suggesting that reference-based objective measures may not be accurate for synthetic signals. The PR system achieves better speech quality than the TTS system and better quality in all three measures than DNN-IRM.

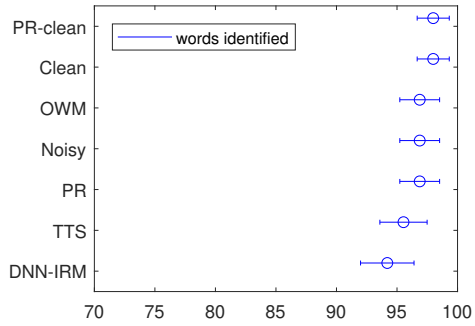

Fig. 2. Subjective intelligibility: percentage of correctly identified words. Error bars show twice the standard error.

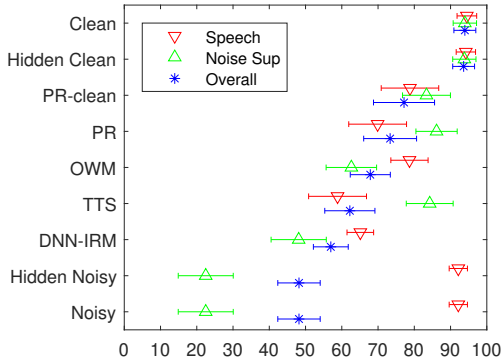

Fig. 3. Subjective quality: higher is better.

5. CONCLUSION

This paper has introduced a speech denoising system inspired by statistical text-to-speech synthesis. The proposed parametric resynthesis system predicts the time-varying acoustic parameters of clean speech directly from noisy speech, and then uses a vocoder to generate the speech waveform. We show that this model outperforms statistical TTS by capturing the prosody of the noisy speech. It provides comparable quality and intelligibility to the oracle Wiener mask by reproducing all parts of the speech signal, even those buried in noise, while still allowing room for improvement as demonstrated by its own oracle upper bound. Future work will explore the extent of speaker-independence that is achievable with this system and other kinds of inputs like filtered and degraded speech [19], and electrophysiological recordings like EEG [20] and ECoG [21].

6. ACKNOWLEDGMENTS

The authors would like to thank Yuxuan Wang for helpful discussions. This material is based upon work supported by the National Science Foundation (NSF) under Grant IIS-1618061. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF.

7. REFERENCES

[1] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, Oct. 2018.

[2] Zhizheng Wu, Oliver Watts, and Simon King, "Merlin: An open source neural network speech synthesis system," *Proc. SSW*, 2016.

[3] Yuxuan Wang, Arun Narayanan, and DeLiang Wang, "On training targets for supervised speech separation," *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1849–1858, 2014.

[4] Hakan Erdogan, John R. Hershey, Shinji Watanabe, and Jonathan Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *Proc. ICASSP*, 2015, vol. 2015-Augus.

[5] Michael I Mandel, Young-Suk Cho, and Yuxuan Wang, "Learning a concatenative resynthesis system for noise suppression," in *Proc. IEEE GlobalSIP Conf*, 2014, pp. 582–586.

[6] Soumi Maiti and Michael I Mandel, "Concatenative resynthesis using twin networks," *Proc. Interspeech*, pp. 3647–3651, 2017.

[7] Soumi Maiti, Joey Ching, and Michael Mandel, "Large vocabulary concatenative resynthesis," in *Proc. Interspeech*, 2018.

[8] Ali Raza Syed, Trinh Viet Anh, and Michael I Mandel, "Concatenative resynthesis with improved training signals for speech enhancement," in *Proc. Interspeech*, 2018.

[9] Dario Rethage, Jordi Pons, and Xavier Serra, "A wavenet for speech denoising," in *Proc. ICASSP*, 2018, pp. 5069–5073.

[10] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W Senior, and Koray Kavukcuoglu, "WaveNet: A generative model for raw audio.," in *Proc. ISCA SSW*, 2016, p. 125.

[11] Masanori Morise, Fumiya Yokomori, and Kenji Ozawa, "WORLD: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Transactions on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.

[12] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[13] John Kominek and Alan W Black, "The CMU arctic speech databases," in *Fifth ISCA workshop on speech synthesis*, 2004.

[14] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third CHiME speech separation and recognition challenge: Dataset, task and baselines," in *Proc. ASRU*, 2015, pp. 504–511.

[15] Diederik P. Kingma and Jimmy Ba, "Adam: A Method for Stochastic Optimization," *arXiv:1412.6980 [cs]*, Dec. 2014.

[16] Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. ICASSP*, 2010, pp. 4214–4217.

[17] Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. ICASSP*. IEEE, 2001, vol. 2, pp. 749–752.

[18] "Method for the subjective assessment of intermediate quality level of audio systems," Tech. Rep. BS.1534-3, International Telecommunication Union Radiocommunication Standardization Sector (ITU-R), 2015.

[19] Michael I Mandel and Young Suk Cho, "Audio super-resolution using concatenative resynthesis," in *Proc. IEEE WASPAA*, 2015.

[20] James A. O'Sullivan, Alan J. Power, Nima Mesgarani, Siddharth Rajaram, John J. Foxe, Barbara G. Shinn-Cunningham, Malcolm Slaney, Shihab A. Shamma, and Edmund C. Lalor, "Attentional selection in a cocktail party environment can be decoded from single-trial EEG," *Cerebral Cortex*, vol. 25, no. 7, pp. 1697–1706, July 2015.

[21] Hassan Akbari, Bahar Khalighinejad, Jose L Herrero, Ashesh D Mehta, and Nima Mesgarani, "Towards reconstructing intelligible speech from the human auditory cortex," *Scientific reports*, vol. 9, no. 1, pp. 874, 2019.