
Proceedings Paper:
Sun, W., Tu, Z. and Ragni, A. orcid.org/0000-0003-0634-4456 (2024) Energy-based models for speech synthesis. In: ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2024), 14-19 Apr 2024, COEX, Seoul, Korea. Institute of Electrical and Electronics Engineers (IEEE) , pp. 12667-12671. ISBN 979-8-3503-4486-8

https://doi.org/10.1109/icassp48485.2024.10447218

ENERGY-BASED MODELS FOR SPEECH SYNTHESIS

Wanli Sun, Zehai Tu, Anton Ragni

Department of Computer Science, University of Sheffield, Sheffield, UK

{wsun20, ztu3, a.ragni}@sheffield.ac.uk

ABSTRACT

Recently there has been a lot of interest in non-autoregressive (non-AR) models for speech synthesis, such as FastSpeech 2 and diffusion models. Unlike AR models, these models do not have autoregressive dependencies among outputs which makes inference efficient. This paper expands the range of available non-AR models with another member called energy-based models (EBMs). The paper describes how noise contrastive estimation, which relies on the comparison between positive and negative samples, can be used to train EBMs. It proposes a number of strategies for generating effective negative samples, including using high-performing AR models. It also describes how sampling from EBMs can be performed using Langevin Markov Chain Monte-Carlo (MCMC). The use of Langevin MCMC enables to draw connections between EBMs and currently popular diffusion models. Experiments on LJSpeech dataset show that the proposed approach offers improvements over Tacotron 2.

***Index Terms*—** speech synthesis, energy-based models, iterative inference

1. INTRODUCTION

Neural network based synthesis have made impressive improvements over statistical speech synthesis. However, these deep learning based text-to-speech (TTS) approaches often feature inconsistencies as did statistical approaches. For example, auto-regressive (AR) models, such as Tacotron 2 [1], Transformer-TTS [2], are almost exclusively trained using teacher forcing [3], where reference rather than predicted values are fed back into the generative process. Such a mismatch between training and inference causes inconsistency called *exposure bias* [4], which may lead to poor generated speech quality (e.g. repetition, skipping, and long pauses [5]). So far there have been a few attempts to alleviate exposure bias, such as, scheduled sampling [6] and attention mechanisms [7]. However, their effective application is complicated due to a number of "training hacks" employed to ensure stable learning [8].

Recently, there has been interest in non-AR models, such as FastSpeech 2 [9] and diffusion models [10]. These models generally do not use teacher forcing as a part of their training and hence should be free of the aforementioned inconsistencies. This paper describes another class of non-AR models called energy-based models (EBMs) [11], which, as will be shown later, have connections to currently popular diffusion models. Given a text, an EBM defines an energy-function over all possible spoken realisations. Although it is possible to formulate the conditional probability distribution of speech given text for EBMs, the intractable normalisation term would make training and inference approaches relying on the probability distribution infeasible.

Instead, training of EBMs can be performed using noise contrastive estimation (NCE), which compares speech data (positive examples), which is assumed to represent high quality speech, and imperfect speech data (negative examples). The nature of imperfection, or negative examples, is crucial when training EBMs. This paper describes a number of effective strategies to generate negative examples, including by means of existing TTS models. Inference with EBMs can be performed using Langevin Markov Chain Monte-Carlo (MCMC) [12, 13]. Given that a similar iterative algorithm is often used with diffusion models (e.g., Grad-TTS [10]), this paper discusses connections between EBMs and diffusion models.

This paper makes the following specific contributions:

1. first energy-based text-to-speech model;
2. a range of methods for generating effective negative samples to use in NCE and elsewhere;
3. link between diffusion models and energy-based models.

The rest of this paper is organized as follows. Section 2 describes energy-based models (EBM), which includes inference, training and negative sampling methods. Section 3 relates EBMs to filtering approaches and diffusion models. Experimental results and discussion are presented in Section 4. Conclusions drawn from this work and future research directions are presented in Section 5.

2. ENERGY-BASED MODELS

Given a text sequence $\boldsymbol{x}$, an energy-based model (EBM) of speech feature sequences $\boldsymbol{Y}$ (e.g. log-Mel spectrograms) can be defined by[1]

$$p_{\boldsymbol{\theta}}(\boldsymbol{Y}|\boldsymbol{x}) = \frac{1}{Z_{\boldsymbol{\theta}}(\boldsymbol{x})} \exp\left(-E_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{Y})\right), \quad (1)$$

where $\boldsymbol{\theta}$ are model parameters, $E_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{Y})$ is an energy between text and speech, $Z_{\boldsymbol{\theta}}(\boldsymbol{x})$ is a normalisation term. Unlike speech signal energies commonly used in models like FastSpeech 2, EBM energies $E_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{Y})$ reflect the correspondence between text $\boldsymbol{x}$ and speech $\boldsymbol{Y}$ pairs. Better matching pairs are expected to yield lower energies and *vice versa*. The normalising term $Z_{\boldsymbol{\theta}}(\boldsymbol{x})$ is intractable to compute exactly. Thus, only certain inference and parameter estimation approaches can be used for EBMs.

2.1. Inference

For tasks where outputs are represented by discrete tokens (*e.g.* characters or words), such as text generation [14] and speech recognition [15], EBMs are commonly used to rerank hypotheses generated during beam search. In contrast, for tasks where outputs are represented by continuous variables, such as speech synthesis, EBMs can

[1] An alternative formulation would involve parameterising the gradient of energy instead. Such an approach is possible due to existence of inference and training approaches that rely only on the gradient of energy.

be used for updating hypotheses themselves. This can be done using Langevin Markov Chain Monte-Carlo (MCMC). The Langevin MCMC is an iterative process where, given an initial hypothesis $\boldsymbol{Y}^{(0)}$, the next hypothesis is obtained by

$$\boldsymbol{Y}^{(N+1)} \leftarrow \boldsymbol{Y}^{(N)} - \lambda \, \nabla_{\boldsymbol{Y}} E_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{Y})|_{\boldsymbol{Y}=\boldsymbol{Y}^{(N)}} + \sqrt{2\lambda}\boldsymbol{Z}^{(N)}, \quad (2)$$

where $\boldsymbol{Z}^{(N)} \sim \mathcal{N}(\boldsymbol{0}, \mu\boldsymbol{I})$, λ is an updating rate and μ is commonly set to 1. The need to specify initial hypotheses $\boldsymbol{Y}^{(0)}$ offers a number of interesting options. In the standard Langevin MCMC initial hypotheses are drawn from a simple prior distribution, such as Gaussian [16]. However, more informative priors, such as high-performing TTS models (e.g. Tacotron 2 and FastSpeech 2), can also be explored.

2.2. Training

Since the normalising factor $Z_{\boldsymbol{\theta}}(\boldsymbol{x})$ is intractable, approaches relying on $p_{\boldsymbol{\theta}}(\boldsymbol{Y}|\boldsymbol{x})$ can not be used with EBMs. Furthermore, popular gradient-based MCMC approaches [17] can not be applied with discrete input, as in this work, and Gibbs sampling [18] would be too computationally expensive. Fortunately, noise contrastive estimation (NCE) [19] provides a feasible solution to optimize energy functions. The NCE loss function for EBMs in eq. (1) is given by

$$\begin{aligned} \mathcal{L}_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{Y}^+, \boldsymbol{Y}^-) = &- \log\left(\frac{1}{1+\exp\left(E_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{Y}^+)\right)}\right) \\ &- \log\left(\frac{1}{1+\exp\left(-E_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{Y}^-)\right)}\right) \end{aligned} \quad (3)$$

where $\boldsymbol{Y}^+$ are called positive samples and $\boldsymbol{Y}^-$ are called negative samples. According to eq. (3), energy functions are optimal when high energy is assigned to negatives and low energy is assigned to positives. Once trained, energy functions can be used for ranking hypotheses generated by other models or inferring hypotheses using the Langevin MCMC in eq. (2).

Positive samples in NCE are usually represented by reference sequences. On the other hand, negative samples need to be designed. In text generation [14] it is argued that negative examples with poor quality make the task of learning the energy function easier which leads to poor quality energy functions. This work proposes using pre-trained TTS models to generate high-quality negative examples. Note that when a pre-trained model is used as a part of training process then it is also possible to adopt it for initialising the Langevin MCMC in eq. 2, which is expected to speed up inference and lead to higher quality hypotheses. The simplest method to generate negative samples from pre-trained TTS models would use hypotheses generated by those models directly. Such approach may fail to work due to high level of similarity between high-quality hypotheses and reference sequences. Other possible options include applying random masking (RM) and SpecAugment [20] (*i.e.* time masking (TM), frequency masking (FM) and time warping (TW)) to those hypotheses. The TM and FM methods can be seen as specific cases of the RM method and may prove less effective. For example, the FM may drastically affect pitch information by masking a whole frequency range. The TW method is also expected to face challenges as utterance-wide shortening/elongation of all sounds may be hard to separate from reference sequences.

2.3. Architecture

The architecture of EBM explored in this work is inspired by Transformer TTS [2] and is shown in Figure 1. The EBM in Fig. 1 consists

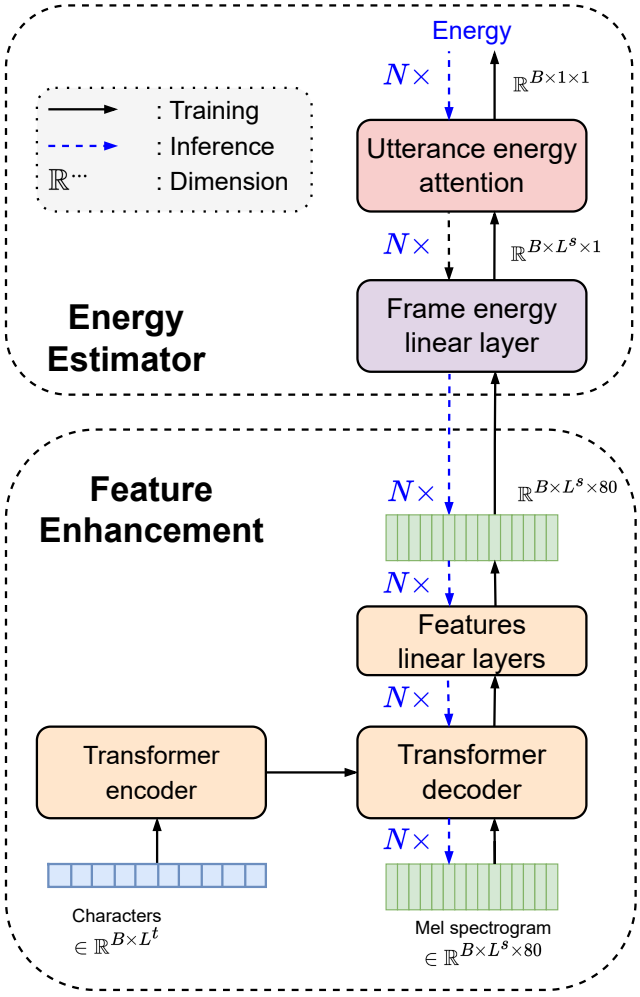

Fig. 1: Architecture of EBMs examined in this work (inspired by Transformer TTS)

of two blocks: energy estimator (top) and feature enhancement (bottom). The goal of the energy estimator is to derive utterance-level EBM energies $E_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{Y})$ from the output of the feature enhancement block. This work assumes that these energies can be derived from frame-level EBM energies. As shown in Fig. 1, the energy estimator consists of two key elements: frame-level EBM energy estimation and frame-level EBM energy weighting. The latter element is motivated by an intuition that frame-level EBM energies are unlikely to make equally important contributions. For example, speech and non-speech frames will likely make different contributions to the utterance-level EBM energy

$$E_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{Y}) = \sum_{t=1}^{T} \alpha_t e_t \quad (4)$$

where $\alpha_{1:T}$ is the sequence of attention weights generated by the EBM energy weighting module, $e_{1:T}$ is the sequence of frame-level EBM energies and T is the number of frames. The attention weights are derived from the frame-level EBM energies. The frame-level EBM energies $e_{1:T}$ are computed by

$$e_t = \boldsymbol{a}^\top \boldsymbol{g}_t + b \quad (5)$$

where $\boldsymbol{g}_t$ is the output of the feature enhancement block, and $\boldsymbol{a}$ and b are parameters of the frame energy module. The goal of the feature enhancement block is to enhance typically short-term spectral information available in standard speech features, such as log-Mel spectrograms, with more advanced acoustic and linguistic information. The estimator is a transformer-based [2] model using text as the input to encoder and and spectral features and the output of encoder as the input to decoder. The output of decoder after linear transformation, $\boldsymbol{g}_t$, is passed to the energy estimator block. Note that the decoder does not use masking to constrain the underlying attention mechanism from attending over previous spectral features, which makes $\boldsymbol{g}_t$ a function of entire text and spectral feature sequences.

3. RELATED WORK

EBMs have been applied in a wider range of domains, e.g. natural language Processing [14, 21] and automatice speech recognition [15]. The EBM proposed in this work can be related to a number of previously proposed approaches in TTS. The use of hypotheses generated by pre-trained TTS models as a part of training and inference allows to connect this EBM to post-filtering methods. Statistical post-filtering approaches, such as [22], aim to address over-smoothing in hypotheses generated by statistical speech synthesis models. However, these approaches suffer from difficulties in accurately modelling probability density functions of the underlying speech parameterisations. Recently, there has also been interest in deep learning based post-filtering approaches. In [23], frequency band specific generative adversarial networks (GAN) were trained to improve the quality of hypotheses generated by deep learning based speech synthesis models. However, this approach assumes independence among frequency bands which may lead to suboptimal results.

More recently, there has been a lot of interest in diffusion-based TTS models [10], which have been extended to audio synthesis [24] and singing voice synthesis [25]. In these models training (forward) and inference (reverse) processes iteratively build a connection between data and noise. Although seemingly different, such diffusion models and the EBM proposed in this work have clear connections. Consider, for example, the iterative inference process used by one of those diffusion models [26]

$$\boldsymbol{Y}^{(N+1)} \leftarrow \frac{1}{\sqrt{1-\lambda_N}}(\boldsymbol{Y}^{(N)} + \lambda_N S_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{Y}^{(N)}, N)) + \sqrt{\lambda_N}\boldsymbol{Z}^{(N)}, \tag{6}$$

Compared to the Langevin MCMC in eq. (2) the key difference stems from modelling iteration, N, specific $S_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{Y}^{(N)}, N)$ score (gradient of log-likelihood) rather than iteration independent $\nabla_{\boldsymbol{Y}^{(N)}} E_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{Y}^{(N)})$ score in the EBM given by eq. (1). In addition, score matching approaches to training diffusion models can also be adopted with EBMs [26], which further strengthens the connection between these models.

4. EXPERIMENTS

4.1. Experimental setup

4.1.1. Dataset

The dataset used in this work is LJSpeech [27], which includes 13,100 audio clips totalling approximately 24 hours from one female speaker. The dataset is split randomly into training (10,000 clips), validation (1800 clips) and test (1300 clips) sets. Objective evaluation is performed over the entire test set whilst subjective evaluation is performed over 100 randomly chosen test set clips. Front-end pre-processing of audio follows the open-source implementation available as a part of NVIDIA's Tacotron 2. [2]

4.1.2. Models

The pre-trained TTS model providing hypothese for training and inference is Tacotron 2 [1]. The open-source implementation of NVIDIA using default configuration was adopted in this work. The structure of EBM follows the corresponding elements of Transformer-TTS [2] available through an open-source implementation [3] except that: 1) positional and character embeddings are 256-dimensional; 2) two EBMs with different dimensions of hidden features, 128 and 256 respectively, are explored in the study. Utterance-level energy is predicted by frame-level EBM energy prediction module, which consists of two 512-dimensional fully-connected layers. Although it is possible to backpropagate gradients through the pre-trained TTS model, for simplicity this was not explored in this work. Both EBMs are trained for 125K iterations using Adam optimizer using batch size of 16 and a constant learning rate of 1×10^{-4} on a single NVIDIA 3090 GPU. The number of parameters of these 2 EBMs and Tacotron 2 are shown in Table 1. We use an open-source implementation[4] of the WaveGlow [28] vocoder and adopt its default settings.

4.1.3. Evaluation

Mel cepstral distortion (MCD) , F0 frame error (FFE) and log-scale F0 root mean square error (log F0 RMSE) are adopted as objective metrics in this work. The MCD metric calculates distance between cepstral coefficient sequences of different lengths on the Mel frequency scale. The FFE metric measures discrepancy of fundamental frequency (F0) between synthesized and reference waveforms. Before objective calculating, dynamic time warping (DTW) is used to align the predicted mel-spectrogram and the reference. FAIRSEQ S^2 toolkit [29] is used to compute MCD and FFE scores. Mean opinion score (MOS) evaluation is conducted to evaluate speech naturalness by scoring each speech sample on a scale between 1 to 5 with 1 point intervals. Waveforms synthesized by 3 models compared in this work are mixed with test set waveforms. Each audio is listened to by 5 listeners, who are native English speakers, on the Amazon Mechanical Turk platform.

Model	**MCD** $\downarrow$	**FFE** $\downarrow$	$\log f_{\text{o}} \downarrow$	**Parameters**
Tacotron 2	4.218	47.31%	0.292	28.19M
EBM$^{(1)}$ (small)	4.163	47.06%	0.289	2.30M
EBM$^{(1)}$ (large)	4.178	47.05%	0.291	7.64M

Table 1: Comparison between Tacotron 2 and two EBMs utilising Tacotron 2 hypotheses as negative samples

4.2. Negative sampling methods

Table 1 compares Tacotron 2 and two initial EBMs. These EBMs were trained using Tacotron 2 generated hypotheses as negative samples and a single step ($N = 1$) Langevin MCMC, where μ was set to 0 for simplicity and Adam rather than gradient descent update rule

[2] https://github.com/NVIDIA/tacotron2

[3] https://github.com/soobinseo/Transformer-TTS

[4] https://github.com/NVIDIA/waveglow

was adopted. Both large (256 hidden features) and small (128 hidden features) EBMs perform slightly better than the baseline.

Table 2 summarises performance of the alternative negative sampling methods (see Sec. 2.2) with the large EBM, where the simplified Langevin MCMC was run for $N = 100$ steps. Many of these methods show significantly better performance than the baseline. Comparing between compressed and stretched spectral features suggest no strong preference for any particular method of time warping (TW). The method of 5% time masking (TM) achieves lower MCD and FFE compared to other time masking methods, while the trend is opposite for frequency masking (FM), where higher percentage points (15%) appear to be yielding better MCD results and worse FFE results. The likely reason is the negative interaction between FFE and frequency masking.

Condition		**MCD** ↓	**FFE** ↓	$\log f_\mathrm{o}$ ↓
TM:	5%	4.149	46.89%	0.290
	10%	4.166	46.98%	0.284
	15%	4.161	47.30%	0.285
FM:	5%	4.166	46.88%	0.292
	10%	4.138	47.27%	0.286
	15%	4.097	47.35%	0.284
TW:	1.2 (compress)	4.134	47.05%	0.291
	1.1 (compress)	4.170	47.03%	0.291
	0.9 (stretch)	4.168	47.28%	0.284
	0.8 (stretch)	4.159	47.29%	0.286
RM:	25%	**3.943**	**46.16%**	0.282
	30%	4.013	46.57%	**0.280**
	Baseline	4.218	47.31%	0.292

Table 2: Negative sampling methods (95% confidence intervals)

Table 3 investigates the impact of the Langevin MCMC steps on the performance of the best system in Table 2. As the number of steps increases, the EBM applying 25% random masking to negative samples performs better and better.

Step	**MCD**↓	**FFE**↓	$\log f_\mathrm{o}$ ↓
0	4.218	47.31%	0.292
1	4.217	47.31%	0.292
100	3.943	46.16%	0.282
300	**3.937**	**45.85%**	**0.276**

Table 3: Simplified Langevin MCMC (95% confidence intervals)

Although random masking appears to be the most effective negative sampling method, the other masking methods may bring additional complementary information. Table 4 summarises performance of different combination approaches involving the RM 30% EBM in Table 2. All combinations examined perform better than using only random masking. The EBM making use of all masking methods performs the best.

4.3. Subjective evaluation

To solicit subjective assessment, a range of listening tests were conducted (see Sec. 4.1.3). Table 5 shows that the proposed EBM shows generally better MOS scores than the baseline Tacotron 2. Furthermore, the detailed breakdown of MOS score counts in Table 6 shows that the EBM significantly reduced the number of MOS scores 2 (-2) and 3 (-27) and increase the number of MOS scores 4 (+26) and 5 (+3).

RM 30%	**TM** 5%	**FM** 5%	**TW** 1.2	**MCD** ↓	**FFE** ↓	$\log f_\mathrm{o}$ ↓
✓				4.013	46.57%	0.280
✓	✓			3.958	46.31%	0.276
✓		✓		3.997	45.63%	0.278
✓			✓	3.927	45.98%	0.267
✓	✓	✓	✓	**3.882**	**45.36%**	**0.258**

Table 4: Combination of negative sampling methods (95% confidence intervals)

Method	**MOS**
Ground Truth	4.53±0.05
Ground Truth (log-Mel + WaveGlow)	4.39±0.07
Tacotron 2 (log-Mel + WaveGlow)	3.77±0.11
EBM (log-Mel + WaveGlow)	3.84±0.13

Table 5: Subjective evaluation

MOS	**1**	**2**	**3**	**4**	**5**
Tacotron 2	0	3	121	344	15
EBM	0	1	94	370	18

Table 6: MOS score counts

5. CONCLUSIONS

This paper proposed a new class of non-autoregressive (non-AR) text-to-speech (TTS) models called energy-based models (EBM). As an example, it shows how powerful forms of EBMs can be designed by adopting architectures of state-of-the-art AR models like Transformer TTS. Although training models like EBMs is more complicated due to the intractability of normalisation terms, a range of training approaches is available. This paper describes how one such approach called noise contrastive estimation (NCE) can be adopted for training. As the NCE critically relies on the quality of negative samples used to contrast reference speech feature sequences, this paper proposed and evaluated a wide range of negative sampling methods. It found that random masking is the single best method but the combination of all proposed methods yielded the best performance. The paper also shows how sampling from EBMs can be performed by means of Langevin Markov Chain Monte-Carlo (MCMC). Since Langevin MCMC is closely linked with an iterative method used by popular diffusion models, the paper discusses similarities between EBMs and diffusion models. The paper concludes by subjective evaluation and finds that the proposed model provides improvements over Tacotron 2. Future work with EBMs will explore score parameterisation and the use of alternative TTS models for architectural and negative sampling choices.

6. REFERENCES

[1] J. Shen et al., "Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions," in *ICASSP*, 2018, pp. 4779–4783.

[2] N. Li et al., "Neural speech synthesis with transformer network," in *AAAI*, 2019, vol. 33, pp. 6706–6713.

[3] R. J. Williams et al., "A learning algorithm for continually running fully recurrent neural networks," *Neural computation*, vol. 1, no. 2, pp. 270–280, 1989.

[4] Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba, "Sequence level training with recurrent neural networks," *CoRR*, vol. abs/1511.06732, 2015.

[5] J. Shen et al., "Non-attentive tacotron: Robust and controllable neural tts synthesis including unsupervised duration modeling," *arXiv preprint arXiv:2010.04301*, 2020.

[6] S. Bengio et al., "Scheduled sampling for sequence prediction with recurrent neural networks," *Advances in neural information processing systems*, vol. 28, 2015.

[7] Mutian He, Yan Deng, and Lei He, "Robust sequence-to-sequence acoustic modeling with stepwise monotonic attention for neural tts," in *Interspeech*, 2019.

[8] E. Battenberg et al., "Location-relative attention mechanisms for robust long-form speech synthesis," in *ICASSP*, 2020, pp. 6194–6198.

[9] Y. Ren et al., "Fastspeech 2: Fast and high-quality end-to-end text to speech," *arXiv preprint arXiv:2006.04558*, 2020.

[10] V. Popov et al., "Grad-tts: A diffusion probabilistic model for text-to-speech," in *ICML*. PMLR, 2021, pp. 8599–8608.

[11] Y. LeCun et al., "A tutorial on energy-based learning," *Predicting structured data*, vol. 1, no. 0, 2006.

[12] Giorgio Parisi, "Correlation functions and computer simulations," *Nuclear Physics B*, vol. 180, no. 3, pp. 378–384, 1981.

[13] U. Grenander et al., "Representations of knowledge in complex systems," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 56, no. 4, pp. 549–581, 1994.

[14] A. Bakhtin et al., "Residual energy-based models for text.," *J. Mach. Learn. Res.*, vol. 22, pp. 40–1, 2021.

[15] Q. Li et al., "Residual energy-based models for end-to-end speech recognition," in *Interspeech*, 2021.

[16] Ferenc Huszár, "How (not) to train your generative model: Scheduled sampling, likelihood, adversary?," *arXiv preprint arXiv:1511.05101*, 2015.

[17] Y. Du et al., "Implicit generation and generalization in energy-based models," *arXiv preprint arXiv:1903.08689*, 2019.

[18] M. Welling et al., "Exponential family harmoniums with an application to information retrieval," *Advances in neural information processing systems*, vol. 17, 2004.

[19] Z. Ma et al., "Noise contrastive estimation and negative sampling for conditional models: Consistency and statistical efficiency," in *CEMNLP*, 2018.

[20] D. Park et al., "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[21] K. Clark et al., "Pre-training transformers as energy-based cloze models," *arXiv preprint arXiv:2012.08561*, 2020.

[22] S. Takamichi et al., "A postfilter to modify the modulation spectrum in hmm-based speech synthesis," *ICASSP*, pp. 290–294, 2014.

[23] T. Kaneko et al., "Generative adversarial network-based postfilter for stft spectrograms.," in *Interspeech*, 2017, pp. 3389–3393.

[24] S. Pascual et al., "Full-band general audio synthesis with score-based diffusion," in *ICASSP*. IEEE, 2023, pp. 1–5.

[25] H. Xue et al., "Learn2sing 2.0: Diffusion and mutual information-based target speaker svs by learning from singing teacher," *arXiv preprint arXiv:2203.16408*, 2022.

[26] Y. Song et al., "Score-based generative modeling through stochastic differential equations," *arXiv preprint arXiv:2011.13456*, 2020.

[27] Keith Ito and Linda Johnson, "The lj speech dataset," `https://keithito.com/LJ-Speech-Dataset/`, 2017.

[28] R. Prenger et al., "Waveglow: A flow-based generative network for speech synthesis," in *ICASSP*. IEEE, 2019, pp. 3617–3621.

[29] C. Wang et al., "fairseq sˆ 2: A scalable and integrable speech synthesis toolkit," *arXiv preprint arXiv:2109.06912*, 2021.