Distillation-free, stable Training of ClariNet Vocoder with Spectral Energy Distance

Eirini Sisamaki
CSD
University of Crete
Heraklion, Greece
sisamaki@csd.uoc.gr

Yannis Pantazis
Institute of Applied & Computational Mathematics, FORTH
Heraklion, Greece
pantazis@iacm.forth.gr

Vassilis Tsiaras
CSD
University of Crete
Heraklion, Greece
tsiaras@csd.uoc.gr

Yannis Stylianou
CSD
University of Crete
Heraklion, Greece
yannis@csd.uoc.gr

***Abstract*—ClariNet provides high-quality speech but it is based on a tricky knowledge-distillation training method with auxiliary losses and large computational requirements. In this work, we apply the Generalized Energy Distance with a repulsive term to train a distillation-free ClariNet Vocoder that offers simplicity, stable and fast training and has low computational requirements. Its theoretical framework is developed, showing that its gradient is stable. Then, we highlight the importance of the repulsive term. Also, possible sources of instabilities during the distillation-based training approaches are presented and ways to ensure stable training are suggested, thus defining a stable distillation-based baseline. Listening experiments using a publicly available database show that the proposed distillation-free ClariNet vocoder outperforms the stable baseline by 34% in terms of MOS, while the perceptual importance of the repulsive term is clearly demonstrated based on ABX preference test.**

***Index Terms*—ClariNet, Generalized Energy Distance, Kullback-Leibler divergence, Parallel-WaveNet, Speech Synthesis, Stable Training**

I. Introduction

Neural-based Text-to-Speech synthesis has made significant progress in recent years, producing high-quality synthetic speech. Consequently, it has been adopted as the mainstream for research and industrial applications. WaveNet [1] is an autoregressive generative model for waveform synthesis that operates at a very high temporal resolution of raw audio samples and produces high-quality audio. However, its sequential generation is too slow for real-time applications. One of the first attempts to accelerate the generation process of WaveNet is by distilling the knowledge of a trained WaveNet into a flow-based model which generates all samples simultaneously. Parallel WaveNet [2] and ClariNet [3] are two prominent examples of knowledge distillation, and they are based on Inverse Autoregressive Flows (IAFs) [4]. IAFs can be regarded as the dual formulation of deep autoregressive modeling in which sampling is performed in parallel. Note that in this case the training procedure requires the estimation of the likelihood which is sequential and slow [2]. To enable feasible training, the knowledge from a pre-trained WaveNet is distilled [5] into a student IAF. However, the two-stage training with distillation is unstable and significantly affects the quality of the synthesized speech [2], [3].

In this work we focus on ClariNet, which is an end-to-end speech synthesis system, where the vocoder is similar to Parallel WaveNet, but with a Gaussian output distribution instead of a mixture of logistics. Since the distribution of the teacher WaveNet is also Gaussian, the Kullback-Leibler (KL) divergence between the output distribution of the teacher and of the student can be expressed in closed form, facilitating the study of the loss function and its behaviour. However, optimizing the KL divergence alone is not sufficient to constrain the student to generate high-quality speech [2], [3]. Additional losses, like power loss measuring spectral distance, aim to address issues like mode collapse [6] and hoarse-sounding speech in the student. A GAN-based loss combined with a single STFT-based auxiliary loss has also been used in the distillation approach [7]. In [8], ClariNet is trained adversarially with Multi-Resolution Spectrogram (MRS) auxiliary loss, using two training approaches.

A stable and consistent method for training implicit generative models was proposed by Gritsenko [9], [11], [13], [17]. Gritsenko applied a spectral (generalized) energy distance (GED) for training a simplified GAN-TTS generator and also used GED in combination with adversarial training. In contrast to spectral losses employed in other recent works [12], [18], GED includes a repulsive term and is the first proper scoring rule applied in the speech synthesis domain.

In this work, we show that another type of parallel generative models of speech (IAFs) may benefit from training with GED. Energy distances have been applied in recent works [21], [22], [24] as a replacement of previously widespread losses, resulting in stable training and alleviating the mode collapse problem of GANs. Their high-quality results have been demonstrated experimentally in image-processing tasks. The effectiveness of energy distances is established on their theoretical properties [23]. Overall, our contributions are summarized as follows: (1) We propose a distillation-free training method for the ClariNet Vocoder with a single loss function: the generalized energy distance (GED). The training is stable and inference produces high-quality generated speech samples. (2) We prove that training with GED has no numerical instabilities. (3) We compare our distillation-free approach with a baseline using a distillation framework. Note that a naive use of distillation framework will result in instabilities in both stages (teacher-student). In this work we create a stable baseline by introducing novel losses in the teacher-student

training. These new loss terms penalize the out-of-range mean and out-of-range scale parameters and are applied to both the teacher and the student, thus enhancing the stability of the training process.

This paper is organized as follows: Section II presents GED, as a single-loss, distillation-free training approach of ClariNet Vocoder. Then, in II-C, we provide the proof of the numerical stability of GED. Section III describes our baseline with our novel loss terms in the distillation framework. Section IV is dedicated to our experiments and their outcomes. Section V concludes the paper.

II. Distillation-free training with GED

A. Background on GED

GED [9] is defined as the energy distance between two multi-resolution spectrograms. GED belongs to a type of loss known as Energy Score. Gneiting and Raftery [11] showed that the repulsive term of GED give rise to a strictly proper scoring rule which in turn implies improved statistical and convergence properties. Gritsenko [9] combined the result of Gneiting and Raftery [11] with the work of Engel et al. [12] to derive his generalised Energy distance based on spectrograms. GED is also related with Maximum Mean Discrepancy (MMD) since a kernel function used in MMD can induce an energy distance [13]. MMD is a popular probability distance measure between two sample distributions. For MMD, the optimization is assumed to be analytically tractable under some specific conditions [14], [15], [16], making MMD methods stable and consistent [17]. However, the proper scoring rule property of GED implies more refined convergence outcomes, relative to the consistency of MMD. Finally, the choice of distance over spectrograms is very fundamental in practice since it emphasizes which features of the generated speech are most important to the human ear [9]. There are recent works that use the Multi-Resolution Spectrogram (MRS) loss for audio synthesis [12], and an MRS-based amplitude distance for speech synthesis [18]. However, one important difference between these approaches and GED, is that they lack the repulsive term. In fact, in the speech synthesis domain, GED is the first loss function of that type which is a proper scoring rule.

B. GED definition

We train our model minimising the Generalized Energy Distance between generated and real data. Let $\{\mathbf{x_i}, \mathbf{c_i}\}_{i=1}^{B}$ be a minibatch of B examples, where each $\mathbf{x_i} \in \mathbb{R}^T$ is a speech segment and $\mathbf{c_i}$ is the corresponding conditioning Mel-spectrogram. The IAF model, f_θ, generates two independent samples $\mathbf{y_i} = f_\theta(\mathbf{z_i}, \mathbf{c_i})$ and $\mathbf{y'_i} = f_\theta(\mathbf{z'_i}, \mathbf{c_i})$ for each conditioning feature $\mathbf{c_i}$, using two independent samples of white noise sequences $\mathbf{z_i}$ and $\mathbf{z'_i}$. The GED minibatch loss is then calculated as

$$L_{GED}(\theta) = \sum_{i=1}^{B} \left(2d(\mathbf{x_i}, \mathbf{y_i}) - d(\mathbf{y_i}, \mathbf{y'_i})\right) \tag{1}$$

with the multi-resolution spectral distance defined as [12]

$$d(\mathbf{x_i}, \mathbf{x_j}) = \sum_{K \in [2^6 \ldots 2^{11}]} \sum_{n} \|s_n^K(\mathbf{x_i}) - s_n^K(\mathbf{x_j})\|_1 + a_K \|\log s_n^K(\mathbf{x_i}) - \log s_n^K(\mathbf{x_j})\|_2 \tag{2}$$

where the first sum is over a geometric sequence of window-lengths from 64 to 2048, while

$$s_n^K(\mathbf{x_i}) = |STFT_n^K(\mathbf{x_i})|^2 \tag{3}$$

denotes the n-th frame index (or time-slice) of the spectrogram of $\mathbf{x_i}$ with window length K. We set the weight $a_K = \sqrt{K/2}$.

1) The crucial role of the repulsive term : The repulsive term renders GED loss to be a proper scoring rule with respect to the distribution over spectrograms of the generated waveform audio. The negative term in (1) $-d(\mathbf{y_i}, \mathbf{y'_i})$ empowers the model to capture the underlying multi-modal conditional waveform distributions of audio given Mel-spectrograms. This repulsive term tries to push the generated samples apart from each other, while the other terms of the loss aim to bring the generated samples close to the real data. Without this term, the quality of generated speech degrades resulting in inferior scores, as it is shown experimentally in [9] and in our work as well.

C. Training with GED is stable

In this section we show that the gradient of GED can be computed efficiently and without numerical instabilities (so that a model can be trained using the Stochastic Gradient Descent method).

Proposition: The partial derivative of GED is bounded, when $s_n^K(\mathbf{y_i})$ is bounded.

Proof: We note that both $\mathbf{y_i} \in \mathbb{R}^T$ and $\mathbf{y'_i} \in \mathbb{R}^T$ are produced by the neural network and have the same distribution. However, the gradients flow only through $\mathbf{y_i}$, while we use stop_gradient at $\mathbf{y'_i}$.

$$s_n^K(\mathbf{x_i}) = \begin{bmatrix} \mathrm{Re}(s_n^K(\mathbf{x_i}))_1^2 + \mathrm{Im}(s_n^K(\mathbf{x_i}))_1^2 \\ \mathrm{Re}(s_n^K(\mathbf{x_i}))_2^2 + \mathrm{Im}(s_n^K(\mathbf{x_i}))_2^2 \\ \vdots \\ \mathrm{Re}(s_n^K(\mathbf{x_i}))_K^2 + \mathrm{Im}(s_n^K(\mathbf{x_i}))_K^2 \end{bmatrix},$$

Replacing $\mathbf{x_i}$ with $\mathbf{y_i}$ in the above formula, we get $s_n^K(\mathbf{y_i})$. We set:

$L_1(i, K, n) = \|s_n^K(\mathbf{x_i}) - s_n^K(\mathbf{y_i})\|_1 =$

$$\sum_{k=1}^{K} |\,\mathrm{Re}(s_n^K(\mathbf{x_i}))_k^2 + \mathrm{Im}(s_n^K(\mathbf{x_i}))_k^2 - \mathrm{Re}(s_n^K(\mathbf{y_i}))_k^2 - \mathrm{Im}(s_n^K(\mathbf{y}_i))_k^2|$$

$$L_2(i, K, n) = \|\log s_n^K(\mathbf{x_i}) - \log s_n^K(\mathbf{y_i})\|_2 = \|\log \tfrac{s_n^K(\mathbf{x_i})}{s_n^K(\mathbf{y_i})}\|_2$$

$$= \sqrt{\sum_{k=1}^{K} \left[\log \frac{\mathrm{Re}(s_n^K(\mathbf{x_i}))_k^2 + \mathrm{Im}(s_n^K(\mathbf{x_i}))_k^2 + \eta}{\mathrm{Re}(s_n^K(\mathbf{y_i}))_k^2 + \mathrm{Im}(s_n^K(\mathbf{y_i}))_k^2 + \eta}\right]^2}$$

where η is a small positive constant to ensure numerical stability. Accordingly we can write $L_3(i, K, n)$ and $L_4(i, K, n)$.

$L_3(i, K, n) = \|s_n^K(\mathbf{y_i}) - s_n^K(\mathbf{y'_i})\|_1$

$L_4(i, K, n) = \|\log s_n^K(\mathbf{y_i}) - \log s_n^K(\mathbf{y'_i})\|_2 = \|\log \frac{s_n^K(\mathbf{y_i})}{s_n^K(\mathbf{y'_i})}\|_2$

1) Computation of the Loss: Combining (1) and (2),

$$L_{GED}(\theta) = \sum_{i=1}^{B} 2d(\mathbf{x_i}, \mathbf{y_i}) - d(\mathbf{y_i}, \mathbf{y_i'}) = \sum_{i=1}^{B} \sum_{K \in \{2^6,\ldots,2^{11}\}} \sum_{n} [2L_1(i,K,n) + 2a_k L_2(i,K,n) - L_3(i,K,n) - a_k L_4(i,K,n)] \quad (4)$$

2) Backward propagation: For model training we need to calculate the gradient vector $\frac{\partial L_{GED}(\theta)}{\partial \mathbf{y_i}} \in \mathbb{R}^T$ and propagate it back to the neural network. Although $L_{GED}(\theta)$ is calculated given complex-valued spectra it turns out that $\frac{\partial L_{GED}(\theta)}{\partial \mathbf{y_i}}$ can be computed using real-valued analysis. Let S be the frame shift and w be a window of size K. Then the elements of the n-th frame of $\mathbf{y_i}$ are $w_m y_i^{((n-1)S+m-1)}$, $m \in \{1,\ldots,K\}$. We know that: $\mathrm{Re}(s_n^K(\mathbf{y_i}))_k =$

$$= \sum_{m=1}^{K} w_m y_i^{((n-1)S+m-1)} \cos \frac{2\pi(k-1)(m-1)}{K}$$

and $\mathrm{Im}(s_n^K(\mathbf{y_i}))_k =$

$$= -\sum_{m=1}^{K} w_m y_i^{((n-1)S+m-1)} \sin \frac{2\pi(k-1)(m-1)}{K}$$

which are real-valued numbers, like $\frac{\partial L_{GED}(\theta)}{\partial \mathrm{Re}(s_n^K(\mathbf{y_i}))_k}$ and $\frac{\partial L_{GED}(\theta)}{\partial \mathrm{Im}(s_n^K(\mathbf{y_i}))_k}$. Hence, we can compute the gradient using the chain rule.

$$\frac{\partial L_{GED}(\theta)}{\partial w_m y_i^{((n-1)S+m-1)}} =$$

$$\sum_{k=1}^{K} \left[\frac{\partial L_{GED}(\theta)}{\partial \mathrm{Re}(s_n^K(\mathbf{y_i}))_k} \frac{\partial \mathrm{Re}(s_n^K(\mathbf{y_i}))_k}{\partial w_m y_i^{((n-1)S+m-1)}} + \frac{\partial L_{GED}(\theta)}{\partial \mathrm{Im}(s_n^K(\mathbf{y_i}))_k} \frac{\partial \mathrm{Im}(s_n^K(\mathbf{y_i}))_k}{\partial w_m y_i^{((n-1)S+m-1)}} \right] = \quad (5)$$

$$\sum_{k=1}^{K} \left[\frac{\partial L_{GED}(\theta)}{\partial \mathrm{Re}(s_n^K(\mathbf{y_i}))_k} \cos \frac{2\pi(k-1)(m-1)}{K} - \frac{\partial L_{GED}(\theta)}{\partial \mathrm{Im}(s_n^K(\mathbf{y_i}))_k} \sin \frac{2\pi(k-1)(m-1)}{K} \right] \quad (6)$$

Now, we need to calculate the corresponding gradient of each term of (4), in order to compute

$$\frac{\partial L_{GED}(\theta)}{\partial \mathrm{Re}(s_n^K(\mathbf{y_i}))_k}$$

$$\frac{\partial L_1(i,K,n)}{\partial \mathrm{Re}(s_n^K(\mathbf{y_i}))_k} = \sum_{k=1}^{K} \mathrm{sign}\left(\mathrm{Re}(s_n^K(\mathbf{y_i}))_k^2 + \mathrm{Im}(s_n^K(\mathbf{y_i}))_k^2 - \mathrm{Re}(s_n^K(\mathbf{x_i}))_k^2 - \mathrm{Im}(s_n^K(\mathbf{x_i}))_k^2\right) \cdot \mathrm{Re}(s_n^K(\mathbf{y_i}))_k^2 \quad (7)$$

and $\frac{\partial L_2(i,K,n)}{\partial \mathrm{Re}(s_n^K(\mathbf{y_i}))_k} = 4 \cdot 0.5 (L_2(i,K,n))^{-1} \cdot$

$$\cdot \sum_{k=1}^{K} \frac{\mathrm{Re}(s_n^K(\mathbf{y_i}))_k}{\mathrm{Re}(s_n^K(\mathbf{y_i}))_k^2 + \mathrm{Im}(s_n^K(\mathbf{y_i}))_k^2 + \eta} \cdot \cdot \log \frac{\mathrm{Re}(s_n^K(\mathbf{y_i}))_k^2 + \mathrm{Im}(s_n^K(\mathbf{y_i}))_k^2 + \eta}{\mathrm{Re}(s_n^K(\mathbf{x_i}))_k^2 + \mathrm{Im}(s_n^K(\mathbf{x_i}))_k^2 + \eta} \quad (8)$$

Without the constant η, the partial derivative $\frac{\partial L_{GED}(\theta)}{\partial \mathrm{Re}(s_n^K(\mathbf{y_i}))_k}$ could become infinite when $\mathrm{Im}(s_n^K(y_i))_k = 0$ and $\mathrm{Re}(s_n^K(\mathbf{y_i}))_k \to 0$. Likewise, we compute $\frac{\partial L_{GED}(\theta)}{\partial \mathrm{Im}(s_n^K(\mathbf{y_i}))_k}$. After the introduction of constant η, the partial derivatives $\frac{\partial L_{GED}(\theta)}{\partial \mathrm{Re}(s_n^K(\mathbf{y_i}))_k}$ and $\frac{\partial L_{GED}(\theta)}{\partial \mathrm{Im}(s_n^K(\mathbf{y_i}))_k}$ are always bounded, and this implies that the partial derivative, $\frac{\partial L_{GED}(\theta)}{\partial w y_i^{((n-1)S+m-1)}}$, is also bounded. That means that the error, that is propagated to update the weights of the neural network, is bounded and that GED does not have numerical instabilities during training.

III. Stable distillation-based baseline

During the conventional distillation training for the ClariNet Vocoder, the teacher predicts Gaussian distribution parameters and the student adapts its weights to approximate the distribution of the teacher. The Kullback-Leibler (KL) divergence measures how two probability distributions differ, and the optimization algorithm tries to minimize this divergence.

A. Kullback-Leibler (KL) divergence

The KL divergence between two distributions q and p of a continuous random variable is:

$$D_{KL}(q||p) := \int_{-\infty}^{+\infty} q(x) \log\left(\frac{q(x)}{p(x)}\right) dx \quad (9)$$

When both $q(x)$ and $p(x)$ are Gaussians, then the above integral has a closed form. Let $q(x) = \mathcal{N}(\mu_q, \sigma_q^2)$ be the student and $p(x) = \mathcal{N}(\mu_p, \sigma_p^2)$ be the teacher distributions, then the KL divergence is [3]:

$$D_{KL}(q||p) = \log \frac{\sigma_p}{\sigma_q} + \frac{\sigma_q^2 - \sigma_p^2 + (\mu_p - \mu_q)^2}{2\sigma_p^2} \quad (10)$$

A well trained teacher WaveNet model has highly peaked output distributions. At the beginning of the training, σ_p and σ_q have very different magnitudes and combining that with potential small σ_p values, numerical problems will occur.

B. Training teacher WaveNet

To ensure stable training of the ClariNet vocoder, it's crucial to address instabilities associated with the WaveNet model. WaveNet, occasionaly, becomes unstable, leading to out-of-range speech samples sounding similar to white noise, known as Type I artifacts [19]. Type I errors stem from training instabilities and are connected to the form of the loss function, in combination with audio segments with small variations around the mean trajectory, such as silences. As both the teacher WaveNet and ClariNet predict Gaussian distribution parameters, we explore methods to train models that mitigate the occurrence of Type I artifacts. The negative log-likelihood

of a Gaussian distribution, $\mathcal{N}(0, \sigma^2)$ and its derivative are respectively

$$L(\sigma) = -\log p(x) = \frac{1}{2}\log 2\pi + \log\sigma + \frac{1}{2}\frac{x^2}{\sigma^2}\ , \qquad (11)$$

$$\frac{\partial L(\sigma)}{\partial \sigma} = \frac{1}{\sigma} - \frac{x^2}{\sigma^3}\ . \qquad (12)$$

In silence segments, it holds that $x \approx 0$ and the $\log\sigma$ term dominates. The derivative of $L(\sigma)$ is positive and pushes σ towards zero. After a silence segment, σ has progressively become very small and if a sample appears with $x \gg \sigma$, then the term $\frac{1}{2}\frac{x^2}{\sigma^3}$ dominates. The derivative of $L(\sigma)$ is negative and pushes σ towards $+\infty$. Then the minimization of $L(\sigma)$ may get stalled in local optima, in short, very far away from the optimal values.

C. Novel loss terms in teacher-student training

Based on above analysis and in order to stabilize the training without modifying the network, we suggest the use of novel additional loss terms which penalize the low or high values of the $\log\sigma$ parameter and the mean, μ, values outside the interval $[-1, 1]$. The out-of-range loss term for $\log\sigma$ is defined as:

$$L_{logScale} = \lambda_1 \max(0, \zeta_1 - \log\sigma) \qquad (13)$$

where we set $\zeta_1 = -7$ thus penalizing scale values below 0.001, while we set $\lambda_1 = 200$, a constant weighting that balances of the loss terms (note, other values for λ_1 are possible). The out-of-range loss for $\log\sigma$ is suggested to be used together with clipping the $\log\sigma$ before the calculation of the KL divergence. The clipping is used to prevent numerical instabilities, while the out-of-range loss tries to correct the weights of the network that have been affected by previous numerical instabilities *by sending a feedback to them*. The out-of-range loss terms for μ are defined as:

$$L_{MeanMin} = \lambda_2 \max(0, -1 - \mu) \qquad (14)$$

$$L_{MeanMax} = \lambda_3 \max(0, \mu - 1) \qquad (15)$$

where we set $\lambda_2 = \lambda_3 = 100$ (these values have been chosen experimentally). These loss terms can be used in the training of the teacher WaveNet and of the student in the ClariNet vocoder. Based on our experiments, these terms stabilize the training even for small batch sizes which are necessary for single GPU training and with noisy databases. Also, they do not affect the speed in the generation stage. On the other hand, they slightly slow down the training.

IV. Experiments

For our experiments, the following loss functions were used for training the ClariNet Vocoder with the default architecture [3]: (1) KL and out-of-range losses (baseline). (2) GED and KL and out-of-range losses. (3) GED only (distillation-free). (4) GED without the repulsive term. Note that the conditions of training as well the quality of the speech database used for training can influence the probability of unstable training. For example, the use of a small batch size (which is suitable for a single GPU training) increases the possibility of unstable training. Note that large batch size, on the other hand, doesn't guarantee a robust training. The batch size was set to 4 and a single GPU (GeForce RTX2080) training has been conducted. All models were trained and tested on a public domain speech dataset (LJSpeech) consisting of 13,100 short audio clips (total length ∼24 hours, sampling frequency 22050Hz) of a single female speaker [10]. LJSpeech contains audio clippings and long silence segments between words. As we have shown, such segments trigger instabilities during training and combined with, the otherwise very good quality of the database, make LJSpeech ideal for our study. For the conditioning Mel spectrogram computation we used 80 dimensions, 1024 fft size, and 256 hop size.

To assess the performance of each method employed in our study (excluding the variant without the repulsive term for a reason explained below), we used the Mean Opinion Score (MOS) (Table I). Our listening test [26], [27] engaged 20 native English speakers. A set of 7 test utterances was synthesized by each model and given to the participants of the listening test. They were asked to evaluate the perceptual sound quality of speech. The possible responses were: 5=Excellent, 4=Good, 3=Fair, 2=Poor, 1=Bad. Participants wore headphones and had no hearing impairments. We also conducted an ABX preference listening test with 10 expert listeners, comparing the model trained with GED and the model without the repulsive term. The latter was not included in the MOS evaluation because it generated samples with a metallic sound, confirming the necessity of the repulsive term. Each one out of 10 sample utterances was synthesized by the model trained with GED and the model without the repulsive term. For each utterance the participants had to choose the best of the two generated samples or neither.

A. Analysis of the results

The MOS and the ABX preference results are shown in Tables I and II, respectively. It is worth mentioning that making use of the novel additional loss terms, no artifacts of type I were detected in our test samples. Based on the MOS results, the best model is the one trained with GED only. This model learned to generate realistic waveform of very good quality *without* the teacher WaveNet. Models utilizing GED generally

TABLE I
MEAN OPINION SCORE (MOS) WITH STANDARD ERROR

Model	MOS
GED + KLDivergence + novel loss terms	3.04±0.16
KLDivergence + novel loss terms (baseline)	3.00±0.15
GED (distillation-free)	**4.03±0.19**
Original Speech	4.99±0.01

TABLE II
RESULTS OF THE ABX PREFERENCE LISTENING TEST.

Model	GED w/o repuls.term	**GED**	Neither
Preference	2%	**96**%	2%

outperformed those that did not. The combination of GED with KL-Divergence resulted in synthesized speech of good quality, in general, but could not surpass the performance of the GED-only model. We note here that in [9] the combination of GED with adversarial techniques has shown improvement upon the GAN-TTS model in terms of MOS.

In the ABX listening test, the participants overwhelmingly preferred the model trained with GED, with a preference rate of 96%, (Table II). It's worth noting that using GED (with the repulsive term) led to a consistent decrease in the validation loss during training. In summary, our experiments show that incorporating GED into the training process yields the best results in terms of speech quality and improving overall performance. A direct comparison between this work and other approaches was not appropriate, because of the lack of authentic implementation of the models used in our experiments.

V. Conclusion

In this work, we propose using Generalized Energy Distance (GED) for training ClariNet Vocoder, simplifying the process by eliminating the need for distillation. This single-stage approach utilizing a single loss function, improves speech quality, ensures stable training, and reduces training time. We also provide theoretical proof of numerical stability with GED. Additionally, we addressed instabilities in our baseline ClariNet Vocoder by introducing novel loss terms that enhance training stability.

Acknowledgment

This work was supported by computational time granted from the National Infrastructures for Research and Technology (GRNET) in the National HPC facility -ARIS- under projects IDs pa210604, and pa211108. YP acknowledges partial support by the Hellenic Foundation for Research and Innovation (H.F.R.I.) under the "Second Call for H.F.R.I. Research Projects to support Faculty members and Researchers" (Project ID: 4753).

References

[1] 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", arXiv, arXiv.org perpetual, non-exclusive license, 2016. doi: 10.48550/ARXIV.1609.03499

[2] A. van den Oord, Y. Li, I. Babuschkin, K. Simonyan, O. Vinyals, K. Kavukcuoglu, G. van den Driessche, E. Lockhart, L.C. Cobo, F. Stimberg, N. Casagrande, D. Grewe, S. Noury, S. Dieleman, E. Elsen, N. Kalchbrenner, H. Zen, A. Graves, H. King, T. Walters, D. Belov and D. Hassabis, "Parallel WaveNet: Fast High-Fidelity Speech Synthesis", arXiv, arXiv.org perpetual, non-exclusive license, 2017. doi: 10.48550/arXiv.1711.10433

[3] W. Ping, K. Peng and J. Chen, "ClariNet: Parallel Wave Generation in End-to-End Text-to-Speech", arXiv, arXiv.org perpetual, non-exclusive license, 2018. doi: 10.48550/ARXIV.1807.07281

[4] D. P. Kingma, T. Salimans, R. Józefowicz, X. Chen, I. Sutskever and M. Welling, "Improving Variational Autoencoders with Inverse Autoregressive Flow", Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems, pp.4736-4744, 2016.

[5] G. E. Hinton, O. Vinyals and J. Dean, "Distilling the Knowledge in a Neural Network", CoRR, vol. abs/1503.02531, 2015, arXiv, arXiv.org perpetual, non-exclusive license. doi: 10.48550/ARXIV.1503.02531

[6] S. Kim, S. Lee, J. Song, J. Kim and S.Yoon, "FloWaveNet : A Generative Flow for Raw Audio", arXiv, arXiv.org perpetual, non-exclusive license, 2018. doi: 10.48550/ARXIV.1811.02155

[7] R. Yamamoto, E. Song and J. Kim, "Probability density distillation with generative adversarial networks for high-quality parallel waveform generation", ArXiv, 2019, vol. abs/1904.04472.

[8] R. Yamamoto and E. Song and J. Kim, "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram", arXiv, arXiv.org perpetual, non-exclusive license, 2019. doi: 10.48550/ARXIV.1910.11480

[9] A. Gritsenko, T. Salimans, R.van den Berg, J. Snoek and N. Kalchbrenner, "A Spectral Energy Distance for Parallel Speech Synthesis", Advances in Neural Information Processing Systems, 13062–13072, vol. 33, 2020.

[10] Keith Ito and Linda Johnson, The LJ Speech Dataset, https://keithito.com/LJ-Speech-Dataset/, 2017.

[11] T. Gneiting and A. E. Raftery, "Strictly Proper Scoring Rules, Prediction, and Estimation", Journal of the American Statistical Association, vol. 102, pp 359-378, 2007.

[12] J. Engel, L.Hantrakul, C.Gu and A.Roberts, "DDSP: Differentiable Digital Signal Processing", arXiv, 2020, arXiv.org perpetual, non-exclusive license. doi: 10.48550/ARXIV.2001.04643

[13] D. Sejdinovic, B. Sriperumbudur, A. Gretton, and K. Fukumizu, "Equivalence of distance-based and RKHS-based statistics in hypothesis testing", Annals of Statistics, vol. 41, pp. 2263-2291, 2013.

[14] "A Kernel Method for the Two-Sample Problem", A. Gretton and K. Borgwardt, M. J. Rasch, B. Scholkopf, A. J. Smola, arXiv, 2008, 0805.2368.

[15] G.K. Dziugaite, D.M Roy and Z. Ghahramani, "Training generative neural networks via Maximum Mean Discrepancy optimization", arXiv, 2015, abs/1505.03906

[16] Y.Li, K. Swersky and R.Zemel, "Generative Moment Matching Networks", arXiv, 2015, arXiv.org perpetual, non-exclusive license. doi: 10.48550/ARXIV.1502.02761

[17] Shen, Cencheng and Vogelstein and Joshua T., "The Exact Equivalence of Distance and Kernel Methods for Hypothesis Testing", AStA Advances in Statistical Analysis, vol.105, 3, pp 385–403, Sep, 2020, doi: 10.1007/s10182-020-00378-1 [Online].

[18] X. Wang, S. Takaki and J. Yamagishi, "Neural source-filter waveform models for statistical parametric speech synthesis", 2019, arXiv.org perpetual, doi: 10.48550/ARXIV.1904.12088

[19] Y.-C. Wu, K. Kobayashi, T. Hayashi, P.L Tobing and T. Toda, "Collapsed Speech Segment Detection and Suppression for WaveNet Vocoder", Interspeech 2018, pp.1988–1992.

[20] S. Takaki, H. Kameoka and J. Yamagishi, "Training a Neural Speech Waveform Model using Spectral Losses of Short-Time Fourier Transform and Continuous Wavelet Transform", arXiv, 2019, arXiv.org perpetual, non-exclusive license. doi: 10.48550/ARXIV.1903.12392 [Online].

[21] M. G. Bellemare, I. Danihelka, W. Dabney, S. Mohamed, B. Lakshminarayanan, S. Hoyer and R. Munos, "The Cramer Distance as a Solution to Biased Wasserstein Gradients", arXiv, 2017, 1705.10743.

[22] "$\alpha-$EGAN: $\alpha-$Energy distance GAN with an early stopping rule", Fangting Ji and Xin Zhang and Junlong Zhao, Computer Vision and Image Understanding, v. 234, p103748, 2023, :https://doi.org/10.1016/j.cviu.2023.103748 [Online].

[23] G.J. Szekely, M.L. Rizzo, (2023). "The Energy of Data and Distance Correlation" (1st ed.). Chapman and Hall/CRC, https://doi.org/10.1201/9780429157158 [Online].

[24] E. Heitz, and T. Chambon, "The Energy Distance as a Replacement for the Reconstruction Loss in Conditional GANs", 2023, January, Journal of Computer Graphics Techniques (JCGT), vol. 12, 1, pp. 29–48, http://jcgt.org/published/0012/01/02/, issn: 2331-7418.

[25] S. Lee, J. Ha, G. Kim, "Harmonizing Maximum Likelihood with GANs for Multimodal Conditional Generation", International Conference on Learning Representations, 2019, https://openreview.net/forum?id=HJxyAjRcFX.

[26] "RECOMMENDATION ITU-R BS.1534-1- Method for the subjective assessment of intermediate quality level of coding systems", 2003, https://api.semanticscholar.org/CorpusID:140119719

[27] "IEEE Recommended Practice for Speech Quality Measurements", IEEE Transactions on Audio and Electroacoustics, 1969, 17, 225-246, https://api.semanticscholar.org/CorpusID:51648844