Adversarial Approximate Inference for Speech to Electroglottograph Conversion
Prathosh A. P. ©, Varun Srivastava ©, and Mayank Mishra ©

Abstract—Speech produced by human vocal apparatus conveys substantial non-semantic information including the gender of the speaker, voice quality, affective state, abnormalities in the vocal apparatus etc. Such information is attributed to the properties of the voice source signal, which is usually estimated from the speech signal. However, most of the source estimation techniques depend heavily on the goodness of the model assumptions and are prone to noise. A popular alternative is to indirectly obtain the source information through the Electroglottographic (EGG) signal that measures the electrical admittance around the vocal folds using dedicated hardware. In this paper, we address the problem of estimating the EGG signal directly from the speech signal, devoid of any hardware. Sampling from the intractable conditional distribution of the EGG signal given the speech signal is accomplished through optimization of an evidence lower bound. This is constructed via minimization of the KL-divergence between the true and the approximated posteriors of a latent variable learned using a deep neural auto-encoder that serves an informative prior. We demonstrate the efficacy of the method at generating the EGG signal by conducting several experiments on datasets comprising multiple speakers, voice qualities, noise settings and speech pathologies. The proposed method is evaluated on many benchmark metrics and is found to agree with the gold standard while proving better than the state-of-the-art algorithms on a few tasks such as epoch extraction.

Index Terms—Speech2EGG, Approximate inference, Adversarial learning, Electroglottograph, epoch extraction, GCI detection.

I. INTRODUCTION

A. Background

HUMAN speech is often said to convey information on several levels [1], broadly categorized as linguistic, para-linguistic and extra-linguistic layers. The semantic and grammatical content of the intended text is encoded in the linguistic layer through the phonetic units. Significant non-lexical information including the speaker-dialect, emotional and affective state is embedded in the para-linguistic layer. The extra-linguistic layer encompasses the physical and physiological characteristics of the speaker and the vocal apparatus. This includes the identity, gender, age, prosody, loudness and other characteristics of the speaker and the physiological status of the vocal apparatus such as the manner of phonation, presence of vocal disorders etc. [2]. The extra-linguistic information is sometimes also referred to as the Voice Quality (VQ), which is primarily characterized by laryngeal and supralaryngeal features arising during phonation [3]. The perceived VQ largely depends upon the phonation, namely the process of converting a quasi-periodic respiratory air-flow into audible speech through the vibrations of vocal folds [4]. Different types of laryngeal functions and configurations give rise to different phonation types. A few examples include breathy voice, falsetto, creaky and pathological voices that refer to the abnormalities in the vocal folds [5].

B. Characterization of Laryngeal Behaviour

Given its wide applicability in speech analysis, it is desirable to characterize laryngeal behaviour. A wide range of methods exist in practice to do so, starting from simple listening tests to invasive optical technique such as laryngeal stroboscopy [6]. One of the most popular class of approaches to characterize laryngeal behaviour is to use speech signal itself to get an estimate of the voice source signal [7]–[13]. This is usually accomplished using the inverse filtering technique (hence the name glottal inverse filtering) under the assumptions specified by the linear source-filter model for speech production [14]. This method is completely non-invasive and requires no additional sensing hardware since it estimates the glottal activity directly from the speech signal. However, these methods heavily depend on the correctness of the model assumptions, accurate estimation of formant frequencies and closed phase bandwidths. This is especially true for high pitched, nasalized and pathological voices. Further the estimated glottal source is known to suffer from the presence of ripples due to improper formant cancellation especially when there is noise in the recording environment or in the voice production mechanism itself [15]. This leads to improper estimation of underlying voice quality measures (Refer Fig. 2 for a depiction this). A popular alternative is electroglottography which is a non-invasive technique to estimate the laryngeal behaviour devoid of the aforementioned limitations of the glottal inverse filtering [16].

C. Laryngography - An Introduction

Laryngography or Electroglottography (EGG) is a technique used to indirectly quantify laryngeal behavior by measuring the
change in electrical impedance across the throat during speaking [17], [18]. A low voltage high frequency current signal is fed through the larynx and the change in the impedance, which depends on the vocal fold contact properties, is recorded [19]. EGG is shown to have multiple correlates of laryngeal behaviour and has proven useful in multiple tasks such as assessment of phonation types [20], [21], gender [22], [23], emotional state and [24], [25], identification of voice pathologies [6].

The amplitude of the EGG waveform is known to vary linearly with the vocal fold contact area [26]. Thus different sections of the EGG mark different laryngeal activities. Further, the EGG waveform is quasi-periodic during the production of voiced phonemes and has zero or very low amplitude during the unvoiced counterparts. The duration between successive positive peaks in the EGG signal correspond to the instantaneous pitch period and are called glottal cycles. For most of the aforementioned use cases, EGG is analyzed cycle-wise and hence parameterized in accordance with a few epochal points within each cycle. The significant quantification measures for the EGG are the instants of glottal opening, glottal closure, location of the maximum glottal opening and start and end of the cycles (Refer Fig. 1 for a sample EGG waveform for two glottal cycles with the corresponding epochs marked). Further, duration of certain events relative to the pitch period such as glottal closure, opening and the skewness of the waveform in every cycle are known to signify several properties of the glottal activity [21], [27]. These are respectively called the contact/closure, open, and speed quotients (Refer Fig. 1). Furthermore, the morphology of the entire EGG waveform over longitudinal cycles serves several clinical applications [28].

D. Need for Estimation of the EGG

Even though EGG offers several advantages over other glottal estimation techniques, additional hardware is required for its extraction which might not be available, especially in non-clinical settings. Further, it is difficult to record EGG for people with thick neck tissues and is sensitive to the sporadic low-frequency body movements [18]. Thus it would be useful if the EGG waveform could be estimated devoid of physical hardware. Many previous works exist to extract information about the parameters of the EGG waveform directly from the speech signal. The typical approach is to estimate the glottal flow waveform and estimate the glottal flow parameters from it [15]. Nevertheless, we hypothesize that the estimation of the EGG waveform (and thus the glottal flow parameters) directly from the acoustic signals is desirable due to the following reasons:

1) Several applications [19], [20], [22] of EGG demand the morphological analysis of the entire waveform rather than the summary parameters. Thus it is desirable to extract the entire EGG signal rather than parameters quantifying it.

2) Even in the case when the application demands only the estimation of the glottal parameters, we hypothesize (and provide empirical evidence) that it is optimal to estimate it from the EGG signal rather than a glottal flow estimate. This is because of the inherent vulnerability associated with the parametric models of glottal flow estimation [29] that suffer heavily if there are deviations from the model assumption, change in recording conditions (see Fig. 2 for an example) and varying voice characteristics (breathy voice, falsetto) [30] which is not present in the EGG waveform.

3) Glottal flow estimators are task-specific and demand a-priori information such as average pitch period [31], glottal closure instants [32], the region of voicing [33] whereas a direct EGG estimator is devoid of all these rather can provide these information as by-product.

4) The availability of the true EGG signals simultaneously acquired with the corresponding speech signal makes it possible to devise an estimator for the EGG without the need for making a model assumption.

5) There are dedicated algorithms proposed in the literature to extract glottal parameters (such as GCI, Open Quotient) from the estimated glottal waveform [34], [35]. However, estimating these from the EGG waveform is much simpler. For example, GCI could be detected trivially with a threshold on the differentiated version of the EGG signal.

E. Contributions

Owing to the aforementioned need and the tremendous success of the data-driven neural models in distributional learning,
we attempt to address the problem of directly estimating the
Electroglottography signal from the speech signal. We approach
the problem from a distribution transformation perspective
and employ the principles of the supervised variational inference
to conditionally generate the EGG signal given a speech segment.
To the best of our knowledge, this is the first attempt to generate
the whole EGG signal from speech signal. We leverage the
abundant availability of the simultaneously recorded speech and
EGG data for this task. If successful, this could replace the
bulky and expensive EGG device and aid in many applications
such as screening of speech disorders, voice quality assessment,
pitch estimation etc. The following list briefly summarizes the
contributions of this work.

1) Formulation of the problem of speech to EGG conversion
from a data-driven distribution transformation perspective.

2) Introduction of a general method of approximate infer-
ence for conditional distribution transformation through
optimization of the evidence lower bound constructed
by minimizing KL divergence between the true and the
approximate posteriors.

3) Use of an informative prior derived from a neural autoen-
coder, that is known to reconstruct the EGG signal.

4) Employing adversarial learning principles for imposition
of the learned informative prior on the latent space of the
distribution transformation network.

5) Demonstration of the efficacy of the method for the speech
to EGG conversion task through rigorous generalization
experiments with several temporal and spectral metrics, on
multiple datasets comprising different speakers, recording
conditions, noise characteristics, voice qualities and
speech pathologies.

The rest of the paper is organized as follows: Sections II-B
and II-C formulate the problem and develop the theory for
the adversarial approximate inference. Section II-D discusses
the nuances of realizing the method using neural networks.
Sections II-E, III-A and III-B describe the experimental proto-
col, data and the metrics used for the assessment. This is followed
by the discussion of results in Section IV and concluding remarks
in Section IV-E.

II. PROPOSED METHODOLOGY

The task of converting speech to EGG is posed as a distribution
transformation problem where the goal is to learn a non-linear
transformation to map samples from the speech distribution to
the EGG distribution. The existence of such a transformation is
motivated by the fact that the underlying physical phenomena
that gives rise to both the speech and the EGG signal is the same.

Concretely, learning the transformation of speech to EGG is
cast as the problem of learning the parameters \( \phi \) that maximize
the conditional probability distribution \( p_{\phi}(Y|X) \) where \( Y \)
and \( X \) represent the EGG and speech signals, respectively.
While maximization of this unknown probability distribution
is an intractable problem in general, recent advances in the
field of deep learning have allowed construction of successful
generative models for estimating and sampling from a
probability distribution, improving upon many of the caveats
presented by classical sampling techniques such as Markov
Chain Monte Carlo (MCMC). A brief outline of two such
methods has been given below.

A. Background on Neural Generative Models

1) Generative Adversarial Networks: The Generative Ad-
versarial Networks (GAN) framework by [36] forms one of
the most popular approaches to deep generative models which
cast distribution learning as a minimax game. Their primary
advantage over classical sampling techniques such as MCMC
lies in their ability for single step generation of samples from
a desired high-dimensional distribution instead of the comput-
tionally intensive repeated sampling in Markov Chains [37].
Optimizing a GAN involves an adversarial game between two
neural networks - a generator \( G(z) \), and a discriminator \( D(x) \),
where the objective is to match the generator distribution \( P_G(x) \)
to the true data distribution \( P_T(x) \). The generator, \( G \) learns an
implicit density by transforming samples from a (known) prior
distribution \( z \sim P(z) \) to samples \( G(z) \) from the generator
distribution, while the discriminator \( D(x) \) predicts the probability
that \( x \) belongs to the true data distribution. The discriminator
acts like a classifier that aims to distinguish between samples
from the true data distribution \( P_T(x) \) and the generator’s distrib-
tion \( P_G(x) \). The game consists of the generator trying to
fool the discriminator into believing that the generator samples
come from the true data distribution, while the discriminator
tries to correctly distinguish between the two. Formally, the
solution to the game is a Nash equilibrium of the following value
function.

\[
\min_{G} \max_{D} V(G, D) = \mathbb{E}_{x \sim P_T(x)} \left[ \log D(x) \right] + \mathbb{E}_{z \sim P(z)} \left[ \log(1 - D(G(z))) \right] \tag{1}
\]

2) Variational Autoencoders: Variational Autoencoders
[38], [39] alongside GANs form the other most popular approach
to deep generative modelling. Inspired by variational bayesian
inference, VAEs form a directed, graphical model where the
distributions of the random variables are parameterized by
neural networks, and the latent variables are assumed to come
from a tractable, explicitly computable density function (such as
the standard normal distribution). The graphical model
consists of a conditional distribution \( p_{\theta}(x|z) \) over the observed
variables, an approximate posterior over the latent variables
\( q_{\phi}(z|x) \) and a specified prior (with known density) \( p(z) \). Then,
it can be shown that [38]:

\[
\log p_{\theta}(x) \geq - D_{KL}[q_{\phi}(z|x)||p_{\theta}(z)] + \mathbb{E}_{q_{\phi}(z|x)}[\log p_{\theta}(x|z)] \tag{2}
\]

where \( \theta, \phi \) are parameters of a neural network. In VAEs, the
equation (2) is used to optimize a lower bound to the likelihood,
since the true likelihood \( p_{\theta}(x) \) which requires marginalization
over the latent variables \( \int_x p_{\theta}(x, z)dz \) is usually intractable.
B. Problem Formulation

The roots of the task of mapping the distribution of speech to EGG lie in learning the conditional distribution of EGG given speech. While GANs have had remarkable success in sampling from an arbitrary data distribution, learning in a conditional setting is known to be unstable and notoriously hard to train [36]. Secondly, GANs in their original formulation cannot incorporate supervised labels (pairs of speech and EGG segments in this case) into the training paradigm. However, several variations of GANs have been proposed both to stabilize the GAN training and impose a conditioning variable into the generative model [40], [41]. On the other hand, variational auto encoders (VAEs) while robust to training variations, make the assumption of putting a standard normal prior on the latent space. This assumption is not necessarily satisfied in practice, especially when the latent distribution is known to deviate from the standard normal distribution (e.g., a multi-modal distribution) [42]. Thus, it is desirable to impose such properties on the distribution of the latent space that aid the process of conditional generation. In this work, we aim to propose a stable conditional generative model that imposes an informative latent prior through adversarial learning, via the principles of the variational inference. Given the immense variability of human utterances in terms of phonemes, co-articulation, speakers, voice types, gender etc., the distribution of speech samples is expected to have a very high entropy. In contrast to this, the distribution of EGG is expected to possess a lower entropy since it embeds very less high entropy. In contrast to this, the distribution of EGG has a lower entropy than speech and thus can be learnt with low learning complexity. Then it is intuitive to map the input speech samples to such a latent space that would reconstruct the EGG well. This can be achieved by minimizing the KL divergence between the EGG conditional distribution, \( p_\theta(z|y) \) and the approximation \( q_\phi(z|x) \) to the true posterior \( p_\theta(z|x) \), that transforms speech to the EGG signal. Mathematically,

\[
D_{KL} [q_\phi(z|x) || p_\theta(z|y)] = \mathbb{E}_{q_\phi}[\log q_\phi(z|x)] - \mathbb{E}_{q_\phi}[\log p_\theta(z|y)]
\]

(3)

Since the distribution over \( p_\theta(y_i) \) does not depend on \( q_\phi(z|x_i) \), we can write the marginal log-likelihood of the EGG distribution as

\[
\log p_\theta(y_i) = D_{KL}[q_\phi(z|x_i) || p_\theta(z|y_i)] + \mathcal{L}(\phi; x_i)
\]

(5)

where the lower bound on the log-likelihood \( \mathcal{L}(\phi; x_i) \) takes the form

\[
\mathcal{L}(\phi; x_i) = -D_{KL}[q_\phi(z|x_i) || p_\theta(z)] + \mathbb{E}_{q_\phi}[\log p_\theta(y_i|z)]
\]

(6)

As \( D_{KL}[q_\phi(z|x_i) || p_\theta(z)] \geq 0 \), we have the following inequality

\[
\log p_\theta(y_i, x_i) \geq \mathcal{L}(\phi; z_i)
\]

(7)

Adopting a maximum likelihood based approach, we optimize this evidence lower bound \( \mathcal{L}(\phi; x_i) \) (ELBO) with respect to the variational parameters \( \phi \) to learn the approximate distribution. If the true posterior distribution were known \( p_\theta(z|y_i) \), then the optimization problem would reduce to \( \max_\phi \mathcal{L}(\phi, \theta; x_i) \). However, in practice, since the true posterior is unknown, it becomes a joint optimization problem over the variational and generative parameters \{ \phi, \theta \} respectively. Hence, the final optimization problem can be stated as

\[
\max_{\phi, \theta} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(\phi, \theta; x_i)
\]

(8)
D. Optimizing the ELBO

While any optimization technique will achieve a lower bound on the log-likelihood of the EGG by optimizing $L$, the quality of the model is critically dependent on the tightness of the variational bound, as well as the assumption that the true data distribution is well approximated by $p_{\theta}(y|z)$.

The parametrization chosen for the inference or recognition model $q_{\phi}(z|x_i)$ naturally decides the tightness of the lower bound derived above. If there exists $\phi^* \in \Phi$ s.t. $q_{\phi^*}(Z|X) = p_{\theta}(Z|X)$ where $\Phi$ is the space of inference parameters, then the ELBO will be a tight bound to $\log p_{\theta}(y_i)$. Unfortunately, this rarely happens in practice, hence, the space $\Phi$ is designed to be as expressive as possible, to allow learning a close approximation. In our model, we parameterize both $q_{\phi}(z|x_i)$, $p_{\theta}(y_i|z)$ as neural networks which are known to be universal function approximators. As pointed out in [43], this is much more efficient than the classical approach of Variational EM maximization where the joint optimization would involve separate parameters for each data point instead of shared global parameters in a neural network.

In the original formulation of Variational Autoencoders [38], [44], the prior $p_{\theta}(z)$ is chosen to be the distribution $\mathcal{N}(0, 1)$ for ease of sampling, tractability and to obtain a closed form expression for the objective function. This limits both the representation capacity of the learnt variational parameters $\phi$ as they are restricted to be close to an arbitrary distribution, and may cause severe underfitting in the worst case. A weak non-representative prior further exacerbates the underfitting problem in a variational setting, as too weak a prior, will lead to very weak variational bound even in the limit of infinite data and perfect optimization [37]. Thus, choosing a good prior is crucial in a variational setting. We address these caveats by building an end to end differentiable model where the distribution $p_{\theta}(z)$ is learnt by a separate autoencoder, known to perfectly reproduce the EGG signal and ensure that the prior over the latent space is informative enough to achieve tight bound on the log-likelihood.

An autoencoder that reconstructs the EGG signal imposes a latent space $z$ and learns a marginal distribution over it given by the following equation

$$p(z) = \int_y p(z|y)p(y)dy$$

(9)

If one assumes a perfect reconstruction of the EGG by the EGG autoencoder (as validated by our experiments), this marginal distribution can be considered a good prior, since it is known to allow EGG reconstruction from the latent space. In all future exposition, we refer to this marginal distribution as $p_{\theta}(z)$, under the assumption that a low empirical reconstruction loss, implies it is close to the optimal prior. Once this prior is learnt through the EGG autoencoder, we enforce learning of the same distribution at the latent layer of the network that converts speech to EGG. This is done by adversarially minimizing the KL divergence $D_{KL}[q_{\phi}(z|x_i) \parallel p_{\theta}(z)]$ term in equation (6). As shown in [45], adversarial training can be used to train $q_{\phi}(z|x_i)$ to be a universal approximation of the posterior, by augmenting the input $x_i$ with random noise $\epsilon$. This allows construction of arbitrary posteriors $q_{\phi}(z|x_i)$, by evaluating the inference model $f_{\theta}(x_i, \epsilon)$ for different values of $\epsilon$. Thus, even with a fixed deterministic mapping from input to the latent space, the posterior will not collapse to a degenerate Dirac delta distribution (i.e. a discontinuous distribution), as the input noise adds a source of stochasticity other than the data generating distribution itself.

Hence, the actual posterior is given by the expression

$$q_{\phi}(z|x_i) = \int_\epsilon q_{\phi}(z|x_i, \epsilon)p_{\theta}(\epsilon)d\epsilon$$

(10)

where $q_{\phi}(z|x_i, \epsilon)$ is the degenerate distribution $\delta(z - f_{\theta}(x_i, \epsilon))$. This fact is corroborated by experiments, where it is seen that inducing noise in the training data generalizes better. In addition, our method removes the normality constraints both on the posterior as well as the prior distribution by using adversarial training to minimize the KL divergence. This would involve a minimax game between the $q_{\phi}(z|x)$ network and a discriminator network $T_{\psi}(z)$, that is poised to detect whether the sample given by the $q_{\phi}(z|x)$ network comes from the prior $p_{\theta}(z)$ (learned through the EGG autoencoder) or not. Mathematically, the following objective function is optimized to minimize the KL term in ELBO.

$$\min_{q_{\phi}} \max_{T_{\psi}} V(q_{\phi}, T_{\psi}) = \mathbb{E}_{z \sim p_{\theta}(z)}[\log T_{\psi}(z)]$$

$$+ \mathbb{E}_{z \sim q_{\phi}(z|x)}[\log(1 - T_{\psi}(z))]$$

(11)

The second term of the ELBO $L(\phi, \theta; x_i)$ (Eq. (6)) is interpreted as the expected EGG reconstruction error at the output given the latent vector, and can be minimized by a number of loss functions. We choose to minimize this error by measuring the cosine distance between the estimated and the ground truth EGG to impose stronger restrictions on the shape of the learnt EGG which is a defining characteristic of the signal. The long-term amplitude of the EGG is known to be an artifact of the measurement apparatus and use of the cosine loss is expected to aid invariance to the superfluous variations in amplitude under different recording and environmental settings. Mathematically, if $\hat{y}$ and $y$ denote the estimated and the true EGG respectively, the cosine distance loss $L(y, \hat{y})$ is given by

$$L(y, \hat{y}) = \cos^{-1}\left(\frac{\langle \hat{y}, y \rangle}{||\hat{y}|| ||y||}\right)$$

(12)

Since we are using (i) the principles of variational inference to optimize an approximation (lower bound) to the true likelihood, (ii) the principles of adversarial training to impose an informative prior on the latent space of the speech to EGG transformer, we name our method Adversarial Approximate Inference (AAI) whose summary is depicted in Fig. 3 and flow-chart in Fig. 4.

E. Implementation Details

Given an input pair of speech and the corresponding EGG signals, we extract data points by framing them into 12 millisecond windows (the method would perform equally well with other window lengths as long as they have one or two pitch periods, on an average) with a stride length of a single sample. During inference, the predictions over overlapping windows are
Fig. 3. Architectural depiction of the proposed method realized using neural networks. The encoder $Q_\phi$ tries to learn the encoding ($z$) from the speech to the EGG space, specified by the EGG encoder $P_{\theta}(z)$. This is facilitated through an adversarial training with the discriminator $T_{\psi}$. Finally, the decoder $P_{\theta}(y|z)$ remaps $z$ to the EGG space to generate the corresponding EGG signal.

Fig. 4. A flowchart of the training and inference procedure on AAI. The training procedure consists of (i) training an EGG autoencoder (shown on the right) and (ii) training a speech to EGG encoder-decoder network where the encoder is adversarially trained to produce representations similar to the representations learnt by the EGG autoencoder. The inference procedure consists of passing a speech sample through the speech to EGG network shown on the left, which outputs the corresponding EGG signal.

averaged and concatenated to obtain the final estimates. All the models including the EGG autoencoder, $q_\phi$ and $p_\theta$ networks are realized by fully connected neural networks of six layers with diminishing neurons interspersed with batch normalization layers. We employ the standard numerical optimization techniques such as stochastic gradient descent to learn all neural network parameters. The EGG autoencoder is first trained independently which is followed by training the $q_\phi$ and $p_\theta$ networks by employing the latent vectors generated by the trained EGG autoencoder. The complete algorithmic procedure of AAI is presented in Algorithm 1. Fig. 5 illustrates the performance of the AAI algorithm on a segment of speech signal. It is seen that the true and the estimated EGGs agree very well with each other both in terms of time and frequency domain characteristics. Fig. 6 depicts a long segment of voiced speech with multiple phonemes with the corresponding true and the estimated EGG signals. It is seen that the true and the estimated EGGs align closely with each other with the corresponding quotients.

III. EXPERIMENTAL SETUP

A. Dataset Description

The effectiveness of the proposed methodology is demonstrated on multiple tasks and datasets with different speakers, voice qualities, multiple languages as well as speech pathologies. All datasets used for learning and evaluation purposes consist of simultaneous recordings of speech and the corresponding EGGs. The generated EGGs are evaluated on several metrics to ascertain their quality, compared to the ground truth EGGs. Parameters of all the networks are learnt on data provided in the book by D.G. Childers [46], referred to as Childers’ data. This
Algorithm 1: Adversarial Approximate Inference (AAI).

Input: Dataset \( D \), Generative Model \( p_\theta(y|z) \), Transformation Model \( q_\phi(z|x) \), Discriminator \( T_\psi(z) \), Prior \( p_\theta(z) \), Noise Distribution \( p_\epsilon(\epsilon) \), Iterations \( K \), Batchsize \( B \).

repeat
for \( k = 1 \) to \( K \) do
 Sample \( \{x^{(1)}, \ldots x^{(B)}\} \) from dataset \( D \)
 Sample \( \{\epsilon^{(1)}, \ldots \epsilon^{(B)}\} \) from \( p_\epsilon(\epsilon) \)
 \( x^{(i)} \leftarrow x^{(i)} + \epsilon^{(i)} \)
 \( z_Q^{(i)} \leftarrow \) Sample from \( q_\phi(z|x^{(i)}) \)
 Compute gradient with respect to \( \theta \)
 \[ g_\theta \leftarrow \frac{1}{B} \sum_{j=1}^{B} \nabla_\theta \left[ \log p_\theta \left( y^{(j)} | z_Q^{(j)} \right) \right] \]
 Compute gradient with respect to \( \phi \)
 \[ g_\phi \leftarrow \frac{1}{B} \sum_{j=1}^{B} \nabla_\phi \left[ - \log \left( 1 - T_\psi \left( z_Q^{(j)} \right) \right) \right. \]
 \[ \left. + \log p_\theta \left( y^{(j)} | z_Q^{(j)} \right) \right] \]
 Update \( \theta, \phi \) using gradients \( g_\theta, g_\phi \)
end for
 Sample \( \{x^{(1)}, \ldots x^{(B)}\} \) from dataset \( D \)
 Sample \( \{\epsilon^{(1)}, \ldots \epsilon^{(B)}\} \) from \( p_\epsilon(\epsilon) \)
 Sample \( \{z_p^{(1)}, \ldots z_p^{(B)}\} \) from \( p_\theta(z) \)
 \( x^{(i)} \leftarrow x^{(i)} + \epsilon^{(i)} \)
 \( z_Q^{(i)} \leftarrow \) Sample from \( q_\phi(z|x^{(i)}) \)
 Compute gradient with respect to \( \psi \)
 \[ g_\psi \leftarrow \frac{1}{B} \sum_{j=1}^{B} \nabla_\psi \left[ + \log \left( 1 - T_\psi \left( z_Q^{(j)} \right) \right) \right. \]
 \[ \left. + \log T_\psi \left( z_Q^{(j)} \right) \right] \]
 Update \( \psi \) using gradients \( g_\psi \)
until convergence of \( \theta, \phi, \psi \)

Fig. 5. Illustration of the AAI algorithm on a segment of a voiced speech with the corresponding Fourier spectra. It is seen that true and the estimated EGGs agree very well with each other both in terms of time and frequency domain characteristics.

Fig. 6. Illustration of the AAI algorithm for speech to EGG conversion: A long segment of voiced speech with multiple phonemes is shown with the corresponding true and the estimated EGG signals. It is seen that the true and the estimated EGGs align closely with each other with the corresponding quotients.

| Table I: Summary of the Datasets Used in the Study. All Models are Learnt on the Childers Dataset and Test on Rest |
|--------------------------------------------------|----------------------------------|-------------------------------|
| Dataset    | No. Utterances | No. Glottal Cycles |
|------------|----------------|-------------------|
| Childers   | 200            | 1339302           |
| CMU        | 3376           | 1033645           |
| VQ         | 120            | 6926930           |
| Pathology  | 1873           | 376905            |

The dataset consists of simultaneous recordings of the speech and the EGG signals from 52 speakers (males and females) recorded in a single wall sound room. Childers’ data consist of utterances of 16 fricatives, 12 vowels, digit counting from one to ten with increasing loudness, three sentences and uttering ‘la’ in a singing voice. For assessing the generalization, the learnt model has been tested on three datasets as described below.

- **CMU ARCTIC** databases [47] consisting of 3 speakers SLT (US female), JMK (Canada male) and BDL (US male) that has phonetically balanced sentences with simultaneously recorded EGGs.
- **Voice Quality (VQ)** database [21] consists of recordings from 20 female and 20 male healthy speakers. The subjects phonated at conversational pitch and loudness on the vowel /a:/ in three ways, (i) habitual voice, (ii) breathy voice and (iii) pressed voice.
- **Saarbruecken Voice** database [48], referred to as Pathology database, a collection of voice recordings from 2000 people both healthy and afflicted with several speech pathologies. The dataset contains recordings of vowels at various pitch levels as well as a recording of the sentence “Guten Morgen, wie geht es Ihnen?” (“Good morning, how are you?”).
B. Evaluation Metrics

Different assessment criteria are for evaluating the performance of the proposed method. As mentioned briefly in the introduction, EGG signal is characterized through multiple metrics that signify the shape of the EGG signal (and its derivative). Most popular metrics that are employed in the literature are the Glottal Closure Instants (GCI), Glottal Opening instants (GOI), Contact Quotient (CQ), Open Quotient (OQ), Speed/Skew Quotient (SQ) and Harmonic to Noise ratio (HNR). All the metrics are computed on the ground truth and the estimated EGGs using the same procedure. In the following section, we discuss each of these metrics in detail (Refer Fig. 1 for a pictorial depiction).

1) GCI Detection: The instant of significant excitation in a single pitch period is defined to be an Epoch which coincides with the instant of maximal closure of the glottis (GCI) [49]. GCIs manifest as significant negative peaks in the differentiated EGG (dEGG) signal (Fig. 1). GCIs are considered to be one of the most important events worthy of detection [34]. Since, our method directly transforms the speech into the corresponding EGG signal, it is devoid of the requirement for the use of auxiliary signals (such as Voice source estimate) to detect GCIs. Further, since GCIs are present only during the voiced speech, most of the state-of-the-art GCI detectors rely on voiced-unvoiced classification as a necessary first step. However, AAI does not demand for a-priori voiced-unvoiced classification as it fully estimates the EGG. We extract GCIs (both true and the estimated EGGs) by picking the negative peaks in the dEGG signal using standard peak-picking algorithms. The de-facto metrics for the GCI detection task, namely, Identification Rate (IDR - % of correct detections), Miss Rate (MR - % of missed detections), False Alarm Rate (FAR - % of false insertions) and Identification Accuracy (IDA - standard deviation of the errors between the predicted and the true GCIs) [34] are used for evaluation and comparison purposes.

2) GOI Detection Task: The complementary task of the GCI detection is the detection of instant of glottal opening. These manifest as positive peaks in dEGG signal (Fig. 1), are usually feeble in magnitude and very susceptible to noise. The same metrics used in the GCI detection problem are used to characterize the performance of GOI detection task as well.

3) CQ: The contact quotient measures the relative contact time or the ratio of the contact time and time period of a single cycle [21].

4) OQ: The open quotient measures the relative open time or the ratio of the period where the glottis is open to the time period of a single cycle.

5) SQ: The speed quotient measures the ratio of the glottal opening time to the glottal closing time, and characterizes the degree of asymmetry in each cycle of the EGG signal. Both the SQ and the CQ have been observed to be sensitive to abnormalities in the glottal flow [50].

6) HNR: The Harmonic to Noise ratio measures the periodicity of the EGG signal by decomposing the EGG signal into a periodic and an additive noise component. It provides a measure of similarity in the spectral domain, by using a frequency domain representation to compare the ratio of energy of the periodic component to the noise component. It is also known to quantify the hoarseness of the voice [51].

The set of quotient metrics, i.e. CQ, OQ, SQ and HNR, both reference and estimated values are computed for every cycle in the ground truth and the estimated EGG and the summary statistics of the true and the estimated values are compared dataset-wise. For GCI detection task, we use the CMU Arctic datasets corrupted with two noise types, stationary white noise and non-stationary Babble noise at five different SNRs till 0 dB at steps of 5 dB. The same model that is trained for clean speech of the Children’s dataset is used for inference in the noise case as well. Further, we compare the proposed algorithm with five baseline state-of-the-art algorithms namely DPI [33], ZFR [31], MMF [52] and SEDREAMS [53].

IV. RESULTS AND DISCUSSION

A. Performance and Comparison

Fig. 8 shows a comparison between AAI and four baseline algorithms - SEDREAMS, DPI, MMF, ZFR on the GCI detection task for two different types of additive noise - synthetic white noise and multiplexer babble noise. We can consistently see superior performance for the AAI algorithm, across all metrics, IDR (higher is better), MR, FAR, IDA (lower is better). Both AAI and ZFR demonstrate robustness to noise even at very low SNRs. The higher performance of our algorithm may be attributed to the fact that we operate directly on the signal of interest (EGG) instead of ancillary signals such as the inverse filtered speech, or its derivatives, which degrade with noise and are severely restricted by the assumptions of the source filter model. In addition, we specifically choose the informative prior $p(z)$ that would encourage learning the same latent representation for both clean and noisy speech, which would help in alleviating the effects of noise. The inverse filtered approach
however displays consistent deterioration with decreasing SNR (e.g., DPI algorithm). The fact that the AAI model offers the best performance despite being trained on clean speech of another dataset, vouches for its generalization capabilities.

Previous work on the task of GCI detection relies on the extraction of voiced region from the EGG signal as a preprocessing step. Our method removes this dependence on the ground truth EGG and accurately captures the voiced and unvoiced regions via a simple energy based method since it precisely estimates the voicing boundaries as shown in Fig. 7.

Tables III, IV, V describe the performance of AAI on different speakers in the CMU dataset, different voice qualities in the VQ dataset and different pathology types in the Pathalogy dataset, respectively. It can be observed that the estimated values of the quotient metrics lie within a small margin of error of the true values. Table III shows significant variation in CQ and SQ for different speaker characteristics, which can be used to distinguish between different speakers. Table IV shows similar variation in CQ and SQ for different voice qualities. The SQ for different pathologies vary significant as seen in V. Thus, any algorithm which achieves a good approximation to these distinguishing characteristics (such as AAI), can be utilized for the purpose for distinguishing on the basis of different speakers, voice qualities, pathology type and so on. While the proposed method is agnostic to the criteria chosen to extract different metrics, our metrics on the VQ dataset are also corroborated by the original work that created the VQ dataset [21]. The estimated HNR for different speakers also well approximates the true HNR with a rank correlation of 1, however the HNR criterion in and of itself is not a distinguishing characteristic as the true values for all speakers are quite similar. In contrast, it has utility in distinguishing between voice qualities [54] where we observe
significant difference between normal and breathy qualities and so is the case with different pathologies as well. Fig. 11 shows the average performance of AAI across all datasets, and the gamut of defined evaluation criteria. Both GCI and GOI have nearly optimal absolute values across all datasets, with nominal decrease in median values on addition of noise. The consistent performance is a testament to the generalization of the network and the efficacy of the method proposed. There is an increase in the variation for GCI and GOI for different voice qualities and pathologies due to the inability to capture certain nuances in the EGG. Figs. 10 and 9 affirm this claim, where AAI is unable to capture the shape at the extrema of the EGG signal. However, it is important to note that the signal characteristics are captured to a significant degree for all the pathologies and voice qualities in which the estimated and the ground truth EGG are virtually indistinguishable unless examined closely.

Table VI experimentally verifies our claim that the cosine loss enforces invariance to amplitude variations while retaining the temporal contours of the generated signal. The quotient metrics CQ, OQ and SQ columns clearly exhibit the inability of the norm based loss ($L_2$) to capture temporal structure accurately, where the cosine distance has much better performance. Even in tasks that depend on amplitude i.e. GCI/GOI detection, the cosine distance has performance equal to or better than the norm loss.

| Feature | Parameters | EGG |
|---------|------------|-----|
| Gender  | 99.59%     | 94.37% |
| Voice Quality | 99.31%  | 92.61% |

To evaluate the meta performance of obtaining the EGG and underscore its utility in various tasks, we demonstrate the classification performance of a shallow fully connected neural network in distinguishing between genders (on CMU) and different voice qualities (on VQ) with a train test split of 70:30 in Table II. The two cases consider either glottal parameters or the EGG itself as input features, and both inputs served as excellent predictors for the tasks outlined. Every input frame is considered as an independent data point for this study. Furthermore, we also computed a pointwise distance to compare the ground truth with the predicted EGG which yielded a $L_2$ norm (averaged across windows) of 0.0045, indicating that the predicted EGG is in fact close to the actual signal.

B. Comparison With Glottal Parameter Estimators

To further test the efficacy of the proposed method, in this section, we compare it with three schemes for glottal inverse
filtering, Iterative Adaptive Inverse Filtering (IAIF) [35], Probabilistic Weighted Linear Prediction (PWLP) [12] and Quasi Closed Phase method (QCP) [10], Conditional Variational Autoencoders (CVAE) and an ordinary multi-layer perceptron regressing on the parameters CQ, SQ and HNR, called Parameter MLP (PMLP). While the inverse filtering based techniques estimate the glottal flow, CV AE and PMLP estimate the EGG and the glottal parameters, respectively. Both QCP and PWLP use weighted linear prediction which is considered more robust to the harmonic structure of speech as compared to standard linear prediction analysis (LP). IAIF uses a multi step iterative sequence of filters to estimate glottal flow and vocal tract envelopes, whose primary advantage is the lack of requirement of any GCI/GOI information. Both QCP and PWLP require GCI/GOI information, however PWLP estimates the closed phase of the source directly from speech instead of GCI location, allowing for better estimation of the voice source. On the other hand in a CVAE [55], a conventional variational autoencoder is built on the EGG signal with an additional conditioning of the speech signal in the latent space that is forced to be a Normal distribution. During inference, only the Decoder is used by conditioning it with the input speech segment to obtain the corresponding EGG signal at the output. The key difference between AAI and a CVAE is that we impose an informative prior (derived from the EGG space) on the latent space through KL minimization achieved via adversarial learning while in a CVAE, the latent space is Normally distributed. Further, since in our model, the aggregated prior is matched (and not the conditional prior unlike the CVAE), the problems associated with the VAE in simultaneously maximizing the likelihood and conditional KL minimization [56] do not exist. These changes, we believe will lead to better generalization as demonstrated through empirical evidence in the subsequent paragraphs. For AAI, CV AE and PMLP, the Childers dataset has been used for training and CMU (with and without noise) and VQ datasets for testing.

Tables VII, VIII, IX compare the performance of AAI against baseline schemes on the CMU dataset, for clean speech and speech corrupted by 0 dB babble and 0 dB white noise respectively.

### TABLE VII

COMPARISON OF AAI WITH SEVERAL GLOTTAL PARAMETER ESTIMATION TECHNIQUES ON CMU DATASET (CLEAN)

|          | GND | AAI | IAIF | PWLP | QCP | CVAE | PMLP |
|----------|-----|-----|------|------|-----|------|------|
| CQ       | 0.42| 0.45| 0.51 | 0.52 | 0.46| 0.28 |
| SQ       | 1.07| 1.07| 1.35 | 1.22 | 1.12| 0.85 | 1.78 |
| HNR      | 2.57| 2.47| 1.92 | 1.41 | 1.31| 2.22 | 2.49 |

### TABLE VIII

COMPARISON OF THE TWO APPROXIMATIONS FOR THE LIKELIHOOD LOSS TERM ON CMU ARCTIC DATASETS

|        | GCI | GOI | CQ   | SQ   | HNR  |
|--------|-----|-----|------|------|------|
| Cosine  | 0.07| 0.63| 0.28 | 0.39 | 0.98 |
| L2      | 94.90| 2.71| 2.39 | 0.38 | 94.81|

### TABLE IX

COMPARISON OF THE TWO APPROXIMATIONS FOR THE LIKELIHOOD LOSS TERM ON CMU ARCTIC DATASETS
The inverse filtering techniques use highly restrictive model assumptions on the voicing process such as assuming the voice filter to be an all pole LTI system. The learning based methods such as CV AE and PMLP relax these model assumptions, however even they are susceptible to the presence of noise in the input signal, as unlike AAI, they do not enforce learning representations invariant to input noise.

It can be seen that AAI matches or outperforms baseline schemes on all cases. All inverse filtering methods have deteriorating performance at higher formants which is reflected in the worse SQ and HNR values. AAI also outperforms the deep learning methods, CV AE and PMLP. PMLP fails to generalise due to the different representations learnt for different kinds of noise corrupted speech, while AAI enforces learning noise invariant representations for speech. This leads to consistently worse performance in CVAE and PMLP parameter estimation when inferring on speech characteristics different from the training set. Table X corroborates this, where the network PMLP trained on Childers data fails to generalise to the VQ dataset, where the speech characteristics are substantially different from the Childer’s data. The consistently superior empirical performance of AAI in parameter estimation across all datasets, further substantiates that the EGG signal is an optimal representation for such tasks as opposed to voice source estimation techniques.

C. Analysis of Latent Representation

Fig. 12 demonstrates the crux of Adversarial Approximate Inference and its effect on the learnt latent representation. The scatter plot in Fig. 12 represents the projection of the learnt latent vectors onto a 2-dimensional plane computed using t-SNE [57]. The first feature to be noted is the distinct clusters formed by the different voice qualities in the VQ Dataset and the CMU dataset, which implies that the distribution over the embeddings of the VQ speech samples \( P(z \mid VQ \text{ speech}) \) and CMU speech samples \( P(z \mid CMU \text{ speech}) \) have different support in the latent space and the network successfully disentangles the factors of variation in the two kinds of voices.

Secondly, the AAI framework motivated by the need for informative priors and robustness to noise in the input signal, enforces learning a representation that discards information that is irrelevant to produce the output. This is demonstrated by the embeddings of clean and noisy speech for the CMU Dataset in Fig. 12 where both clean and noisy are inseparable in the latent space and the projections. It is desirous to learn a similar latent representation for both noisy and clean speech, since the output of the model i.e. the laryngograph signal remains the same in both scenarios. Both these features aptly demonstrate the efficacy of our technique in mitigating the effects of noise by discarding noise in the input signal, while successfully retaining information that can distinguish between different auditory colorings (or voice qualities).

D. Limitations

In order to evaluate its robustness, AAI’s efficacy was demonstrated across a number of tasks and across the voice diapora. However, being a non linear processing system, the method is still susceptible to the problem of incorrect speech polarity. The asymmetry of the glottal waveform implies that the speech signal is also asymmetric [58], which manifests as different speech polarities when measuring via a microphone. Fig. 13 demonstrates this issue, where we have inverted polarities for speech and the estimated EGG, as our method assumes a positive polarity for the speech signal, defined in [58]. Where applicable, we have manually corrected for such artifacts by inverting the speech polarity. Part (b) in Fig. 13 displays a segment of a pathology...
where AAI fails to capture the higher order harmonics. For robustness to noise, our method utilizes singly strided frames across the speech signal, which consequently acts as a smoothing operation and severely attenuates higher order harmonics. This can be a potential limitation when working with voices which inherently have significant amplitude at high harmonics. However, the method can be made to capture such signal behaviours by having non-overlapping windows.

### E. Conclusion

We proposed a distribution transformation framework to map speech to the corresponding EGG signal, that is robust to noise, and generalizes across recording conditions, speech pathologies and voice qualities. In essence, AAI is a unifying framework for the complete class of methods that create task specific representations and techniques for exploiting the information available in the electroglottographic signal. While the efficacy of AAI is empirically verified in the setting of speech transformation, the constructed framework is agnostic to the application chosen, and can be used in an array of problems. Since learning conditional distributions is a task of ubiquitous importance, future work on this framework can focus on other areas in which similar principles may be applied, and further investigate the statistical properties of the latent space constructed by our model.

### ACKNOWLEDGMENT

The authors would like to thank Dr. T. V Ananthapadmanabha for his valuable comments on the manuscript, Prof. Prasanta Kumar Ghosh for lending the Childer’s dataset, Prof. Anne-Maria Laukkanen for sharing the VQ dataset. They would also like to thank the authors of IAIF, QCP, PWLP and CVAE for providing the codes for their implementations. They would thank the associate editor and the reviewers who helped to enhance the quality of the work significantly.

### REFERENCES

[1] J. Laver and L. John, *Principles of Phonetics*. Cambridge, U.K.: Cambridge Univ. Press, 1994.

[2] K. Marasek, “EGG and voice quality.” Univ. Stuttgart, Stuttgart, Germany, 1997. [Online]. Available: http://www.ims.uni-stuttgart.de/phonetik/EGG/frmst1.htm

[3] R. L. Trask, *A Dictionary of Phonetics and Phonology*. Abington, U.K.: Routledge, 2004.

[4] P. Ladefoged, I. Maddieson, and M. Jackson, “Investigating phonation types in different languages,” in Vocal Physiology: Voice Production, Mechanisms and Functions, O. Fujimura, Ed. New York, NY, USA: Raven Press, 1988, pp. 297–317.

[5] M. Hirano, *Clinical Examination of Voice*. New York, NY, USA: Springer-Verlag, 1981.

[6] P. Kiriz, “Stroboscopy—a pertinent laryngological examination...” *J. Otolaryngol.*, vol. 14, no. 3, pp. 151–157, 1985.

[7] C. Gobl and A. N. Chasaide, “Acoustic characteristics of voice quality,” *Speech Commun.*, vol. 11, no. 4–5, pp. 481–490, 1992.

[8] D. Wong, J. Markel, and A. Gray, “Least squares glottal inverse filtering from the acoustic speech waveform,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 4, pp. 350–355, Aug. 1979.

[9] H. Veeneman and S. BeMent, “Automatic glottal inverse filtering from the acoustic speech waveform,” *IEEE Trans. Audio Electroacoust.*, vol. AE-21, no. 5, pp. 417–427, Oct. 1973.

[10] M. Airaksinen, T. Raitio, B. Story, and P. Alku, “Quasi closed phase glottal inverse filtering analysis with weighted linear prediction,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 3, pp. 596–607, Mar. 2014.

[11] M. Airaksinen, T. Bäckström, and P. Alku, “Quadratic programming approach to glottal inverse filtering by joint norm-1 and norm-2 optimization,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 5, pp. 929–939, May 2017.

[12] A. Rao and P. K. Ghosh, “Glottal inverse filtering using probabilistic weighted linear prediction,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 1, pp. 114–124, Jan. 2019.

[13] T. Raitio et al., “HMM-based speech synthesis utilizing glottal inverse filtering,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 1, pp. 153–165, Jan. 2011.

[14] J. Makhoul, “Linear prediction: A tutorial review,” *Proc. IEEE*, vol. 63, no. 4, pp. 561–580, Apr. 1975.

[15] P. Alku, “Glottal inverse filtering analysis of human voice production – A review of estimation and parameterization methods of the glottal excitation and their applications,” *Sadhana*, vol. 36, no. 5, pp. 623–650, 2011.

[16] D. Veeneman and S. BeMent, “Automatic glottal inverse filtering from speech and electroglottographic signals,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-33, no. 2, pp. 369–377, Apr. 1985.

[17] R. J. Baken, “Electroglottography,” *J. Voice*, vol. 6, no. 2, pp. 98–110, 1992.

[18] D. G. Childers and A. K. Krishnamurthy, “A critical review of electroglottography,” *Cirt. Rev. Biomed. Eng.*, vol. 12, no. 2, pp. 131–161, 1985.

[19] I. R. Titze, “Interpretation of the electroglottographic signal,” *J. Voice*, vol. 4, no. 1, pp. 1–9, 1990.

[20] K. L. Peterson, K. Verdolini-Marston, J. M. Barkmeier, and H. T. Hoffman, “Comparison of aerodynamic and electroglottographic parameters in evaluating clinically relevant voicing patterns,” *Ann. Otol., Rhinology, Laryngology*, vol. 103, no. 5, pp. 335–346, 1994.

[21] D. Liu, E. Kankare, A.-M. Laukkanen, and P. Alku, “Comparison of parameterization methods of electroglottographic and inverse filtered acoustic speech pressure signals in distinguishing between phonation types,” *Biomed. Signal Process. Control*, vol. 36, pp. 183–193, 2017.

[22] Y. Chen, M. P. Robb, and H. R. Gilbert, “Electroglottographic evaluation of gender and vowel effects during modal and vocal fry phonation,” *J. Speech, Lang., Hearing Res.*, vol. 45, no. 5, pp. 821–829, 2002.

[23] M. B. Higgins and L. Schulte, “Gender differences in vocal fold contact computed from electroglottographic signals: the influence of measurement criteria,” *J. Acoustical Soc. America*, vol. 111, no. 4, pp. 1865–1871, 2002.

[24] T. Waaramaa and E. Kankare, “Acoustic and egg analyses of emotional utterances,” *Logopedia Phoniatrics Vocology*, vol. 38, no. 1, pp. 11–18, 2013.

[25] P. J. Murphy and A.-M. Laukkanen, “Electroglottogram analysis of emotionally styled phonation,” in *Multimodal Signals: Cognitive and Algorithmic Issues*, A. Esposito et al., Eds. Berlin, Germany: Springer, 2009, pp. 264–270.
