Effects of Sinusoidal Model on Non-Parallel Voice Conversion with Adversarial Learning

Mohammed Salah Al-Radhi, Tamás Gábor Csapó and Géza Németh

1 Department of Telecommunications and Media Informatics, Budapest University of Technology and Economics, 1111 Budapest, Hungary; csapot@tmit.bme.hu (T.G.C.); nemeth@tmit.bme.hu (G.N.)
2 MTA-ELTE Lendület Lingual Articulation Research Group, Hungarian Academy of Sciences, 1088 Budapest, Hungary
* Correspondence: malradhi@tmit.bme.hu

Featured Application: The work discussed herein gives an advanced novel approach to improve voice conversion performance that does not require a parallel corpus for training. It can be applied in virtual–augmented reality systems (voice avatars) and other speech assistance systems for overcoming speech impairments.

Abstract: Voice conversion (VC) transforms the speaking style of a source speaker to the speaking style of a target speaker by keeping linguistic information unchanged. Traditional VC techniques rely on parallel recordings of multiple speakers uttering the same sentences. Earlier approaches mainly find a mapping between the given source–target speakers, which contain pairs of similar utterances spoken by different speakers. However, parallel data are computationally expensive and difficult to collect. Non-parallel VC remains an interesting but challenging speech processing task. To address this limitation, we propose a method that allows a non-parallel many-to-many voice conversion by using a generative adversarial network. To the best of the authors’ knowledge, our study is the first one that employs a sinusoidal model with continuous parameters to generate converted speech signals. Our method involves only several minutes of training examples without parallel utterances or time alignment procedures, where the source–target speakers are entirely unseen by the training dataset. Moreover, empirical study is carried out on the publicly available CSTR VCTK corpus. Our conclusions indicate that the proposed method reached the state-of-the-art results in speaker similarity to the utterance produced by the target speaker, while suggesting important structural ones to be further analyzed by experts.

Keywords: sinusoidal model; non-parallel voice conversion; generative adversarial networks; continuous parameters

1. Introduction

Voice conversion (VC) is a task developed to convert the observed identity of a source speaker to sound like a different target speaker, while retaining the linguistic or phonetic content unchanged. This is achieved through adjusting the spectral and prosodic features of the input speaker [1]. This technology has been applied to many potential tasks, such as speech synthesis [2], speech enhancement [3,4], normalization of impaired speech [5,6], and singing style conversion [7] and can also be used for generating new voices for animated and fictional movies. Towards the practical use of these applications, it is necessary to modify the VC approach.

A large number of popular approaches easily construct a conversion function with either a statistical Gaussian mixture model (GMM) [8] or Gaussian process regression (GPR) [9] that modifies acoustic features (such as mel-frequency cepstral coefficients) between source and target speakers. However, the simplicity of these models and the lack
of vocoding algorithms are prone to generating robotic-sounding results. Several non-linear spectral mapping techniques address these constraints: frequency warping [10], non-negative matrix factorization (NMF) [11], restricted Boltzmann machine (RBM) [12], artificial neural networks (ANNs) [13], feed-forward deep neural networks (FF-DNNs) [14], and long short-term memory (LSTM) [15]. With the novelty of WaveNet architecture, it was applied in [16] as a vocoder and the quality of synthesized speech was greatly increased. Nevertheless, most of these conventional VC methods require accurately aligned parallel data for training, where the source speaker and target speaker should read the same sentence, and their pairs of frames must be aligned. On the other hand, it is not always feasible to collect parallel utterances. Even when parallel data are accessible, the required alignment procedures introduce artifacts and lead to speech-quality degradation. Furthermore, these methods may need some sort of automatic time alignment between the speech data of speakers if they were not accurately time aligned. This can lead to a trick and influence the training part of the VC task [17]. In addition, there is another major problem: if both source and target speakers speak different languages or have different accents, building parallel data would become an unviable task.

To overcome these limitations, numerous attempts have been made to use non-parallel voice conversion methods. The concept of this paradigm is that voice examples of many speakers are provided, but these examples do not appear in the same sentence. In other words, linguistic features are not shared between speakers. Hence, the VC model can generalize to different voices in the same dataset or even outside it. This type of voice conversion technique can be reliably applied as one-to-many [18] and many-to-many VC [19]. The converted speech quality of non-parallel VC systems is still likely not better than that available from systems using parallel data. Thus, suggesting a non-parallel VC technique can be very challenging to estimate natural voices.

Over the years, a well-designed VC system has typically consisted of an analysis of the input waveform into acoustic features, a conversion function to map the source–target features, and synthesizing the converted features to a target speaker. Note that these phases may degrade the performance of the entire voice conversion system. For this, the speech analysis and synthesis system (also referred to as vocoder) is of principal significance. Many attempts have been made previously to construct a parametric vocoder to model voice signals. We can group the vocoders based on their implementations into three main categories: source-filter [20], phase [21], and neural complex [22] models. However, most of the recent VC literature has focused only on source-filter and neural models. This narrow focus and lack of phase approach has caused important gaps in the literature.

Sinusoidal modeling is another signal processing technique that offers promise for improving voice system performance. It can be characterized by the amplitudes, frequencies, and phases of the component sine waves. It is also synthesized as the sum of several sinusoids that can generate high-quality speech. Concisely, voiced speech can be modeled as a sum of harmonics (quasi-periodic) with instantaneous phases. On the other hand, unvoiced speech, which is non-periodic, can be formed only by white noise (which means that the spectrum of the noise-like excitation is flat and the number of harmonics is 0). To the best of the authors’ knowledge, a sinusoidal-based phase model has not been investigated yet in non-parallel VC. To address this challenge, hence, this work seeks to fill this gap in the literature by conducting the first study to accurately determine the effect of the sinusoidal model in non-parallel VC, which does not include parallel utterances or time alignment.

The rest of this paper is structured as follows: Section 2 briefly introduces a related work that motivates our study. Section 3 gives details of the continuous sinusoidal model we used for speech synthesis, and Section 4 explains the novel method we used for non-parallel voice conversion. Section 5 discusses the experimental setup and evaluation results. Key points and future work are finally given in Section 6.
2. Related Work and Motivation

In recent times, successful efforts have been considered to develop non-parallel methods. Among them, a conditional variational autoencoder (CVAE) framework was proposed in [23], consisting of two main networks: encoder and decoder. The idea is that the input speech samples are firstly converted by an encoder to latent vectors having the input linguistic variables. Then, a one-hot vector having the target speaker identity is created and later fed together with latent vectors into a decoder to produce the target utterance features. This approach is called a conditional VAE due to the decoder which is trained on a target speaker identity vector. One of the problems encountered here is that the decoder is over-smoothed. This over-smoothing effect is caused by the Gaussian assumptions on the encoder and decoder distributions. In addition, over-smoothness mostly occurs in spectral features as they have low-dimensional features. Parametric techniques usually suffer from over-smoothing because they use the minimum mean square error or the maximum likelihood function as the optimization criterion. Consequently, VAE fails to capture the desired details of temporal and spectral dynamics. This commonly results in poor-quality converted speech.

One powerful technique that can possibly overcome the weakness of CVAEs is the cycle-consistent generative adversarial network (CycleGAN), which was proposed in [24] using generators, discriminators, and an identity-mapping loss. The idea is that the source features are transformed to match the features of a target voice via a GAN model. Then, the result is again converted back to match the source characteristics via an additional GAN. A conjunction of both cycle consistency and adversarial losses is finally used to force the linguistic contexts to be retained in the converted speech. The adversarial loss only tells us whether the generator follows the distribution of target data and does not ensure that the contextual information is preserved. The cycle-consistency loss is introduced to encourage the generator to find the mapping that preserves underlying linguistic content between the input and output. For these reasons, CycleGAN represents a successful deep learning implementation to find an optimal pseudo pair from non-parallel data of paired speakers. It does not require any frame alignment mechanism such as dynamic time warping or attention. While this approach was found to work reasonably well in one-to-one mappings, CycleGAN fails to work in many-to-many VC tasks because it may require several generator–discriminator pairs trained separately. Hence, this increases the computational complexity of learning parameters, training time, and the memory space prohibitively. For VC applications, this can degrade conversion performance as multiple speakers are trained simultaneously. To resolve this issue, StarGAN-VC [25] was recently introduced as a unified model architecture. It accepts concurrent training of various domains, i.e., many-to-many mapping within a single network. It can also be noted that StarGAN gives good quality among the GAN-based VC frameworks. However, it lacks stable training [26], which can be overcome by the proposed sinusoidal model using continuous parameters.

Despite the vast amount of research that has been provided in the literature for non-parallel voice conversion, the synthesized speech obtained from converted features is still far from achieving the quality of the target speaker. Therefore, the new generation of high-quality converted speech is still very challenging and has room for improvement. Motivated by this, we propose a new less expensive sinusoidal architecture as an independent study, which is the first GAN-based sinusoidal model to perform non-parallel VC with only a few training examples. We observe that the proposed model is able to give objective results more efficiently than the StarGAN scheme, and can achieve a good speech quality to the target speaker.
3. Proposed Continuous Sinusoidal Model

This section will introduce our developed sinusoidal model, a framework to address a speech analysis–synthesis system derived from continuous parameters.

Based on a sinusoidal model, an analysis/synthesis system is characterized in terms of the amplitudes, frequencies, and phases of the component sine waves to synthesize high-quality speech. The scope of this scenario is designed to model the voiced speech as a sum of quasi-harmonics with instantaneous phases and fundamental frequency (F0) [27]. Conventionally, the F0 contour is discontinuous at voiced–unvoiced (V-UV) boundaries because F0 is not classified in unvoiced sounds. This can cause some issues in statistical modeling, which involves building separate models for voiced and unvoiced frames of speech. On the other hand, the continuous F0 does not need modeling around V-UV and UV-V transitions. In this study, the V/UV decision can be left up to the maximum voiced frequency feature. Moreover, recent work has demonstrated that a neural vocoder (such as the WaveNet model) yields state-of-the-art performance and gives good-sounding speech. However, it requires a large quantity of data, computation power, and it needs to be repeated sequentially (one sample at a time), making it difficult to train for real-time implementation, especially in embedded environments. Therefore, a continuous sinusoidal model (CSM) is built to tackle the limitations of discontinuity in the speech parameters and the complexity of neural vocoders. Unlike the standard source-filter vocoders, CSM uses harmonic features to simplify and enhance the synthesizing phase prior to reconstruction. Figure 1 shows the main components of the developed sinusoidal model.

![Figure 1. Overview of the developed system. CSM consists of three analysis algorithms (for determining the contF0, spectral envelope, and MVF) and a synthesis algorithm incorporating these parameters.](image)

3.1. Analysis Phase: Feature Extraction

Here, we have designed our continuous sinusoidal model (i.e., in which all parameters are continuous) using three acoustic parameters (contF0, maximum voiced frequency and mel-generalized cepstrum). We use a continuous F0 (not a discontinuous F0 such as Swipe, Yin, DIO, etc.) tracker which takes non-zero pitch values even when voicing is not present and does not apply a strict voiced/unvoiced decision. We choose the continuous modeling proposed in [28] as it can be more effective in achieving natural synthesized speech. Another excitation continuous parameter is the maximum voiced frequency (MVF) [29], which was recently shown to result in a major improvement in the quality of synthesized speech. During the synthesis of various sounds, the MVF parameter can be used as a boundary frequency to separate the voiced and unvoiced components. Therefore, contF0, MVF, and MGC [30] parameter streams are calculated during the analysis phase to be converted. The advantage of this CSM is that it is relatively simple; it has only two one-dimensional parameters for modeling excitation (contF0 and MVF) and the synthesis part is computationally feasible, therefore speech generation can be performed in real time.

It should be noted that contF0 is still sensitive to additive noise in speech signals and suffers from short-term errors (when it changes rather quickly over time). Moreover, it can cause some tracking errors when the speech signal amplitude is low, voice is creaky, or there is a low harmonic-to-noise ratio (HNR). To mitigate these issues, an instantaneous frequency-based (i.e., the time-derivative of the phase) method is employed. In our im-
Implementation, the formula for computing the instantaneous frequency $IF(\tau)$ is given by Flanagan’s equation [31]:

$$IF(\tau) = \frac{a db dt - b da dt}{a^2 + b^2}$$

(1)

where $a$ and $b$ are, respectively, the real and imaginary parts of the spectrum of a waveform $S(\omega)$. Consequently, $contF0$ can be further corrected by

$$Corrected \ contF0 = \frac{\sum_{k=1}^{K} |S(k\omega_0, t)| \cdot IF(k\omega_0, t)}{\sum_{k=1}^{K} k|S(k\omega_0, t)|}$$

(2)

where $K$ represents the harmonics number used for refining (we set $K = 6$), $\omega_0$ denotes the angular frequency of the $contF0$ at a temporal position $t$. The new corrected $contF0$ performance can be compared with the performance of a reference pitch contour ($f0_egg$, estimated from the electroglottograph) as shown in Figure 2. We can see that there is almost equivalent performance between the refined $contF0$ and the $f0_egg$, and much better than the baseline [28]. It can be also observed in the unvoiced region (frames from 170 to 202 in Figure 2) that the new $contF0$ is reduced more significantly than for the baseline.

![Figure 2. Examples of F0 contours from a female speaker extracted by the baseline, EGG, and corrected methods.](image)

In general, the spectral envelope must be characterized using a small number of parameters, either for reaching a high compression rate or for statistical modeling. Preferably, a mel-generalized cepstral (MGC) analysis is implemented on the speech signals, as these features have shown their efficiency to accurately and robustly capture the spectral envelope. In this study, we followed the CheapTrick algorithm proposed by Masanori [30]. Thus, we found 36 coefficients are enough to run our synthesized converted speech. As a result, Figure 3 shows three streams of acoustic information extracted from a speech frame: F0, MVF, and the spectral envelope. We then had an acoustic vector with 36 MGCs, 1 $contF0$, 1 MVF.
3.2. Synthesis Phase: Synthesized Speech

The synthesis algorithm applied in CSM constructs the speech frames into a voiced component $s_v(t)$ and noise component $s_n(t)$ in accordance with MVF values. This can be described by

$$s(t) = s_v(t) + s_n(t)$$  \hspace{1cm} (3)

In voiced frames, the harmonic part can be calculated by the following general formula as:

$$s_v^i(t) = \sum_{k=1}^{K_i} A_i^k \cos\left( k w_0^i t + \phi_i^k \right)$$  \hspace{1cm} (4)

$$w_0^i = \frac{2\pi \text{contF0}^i}{F_s}$$  \hspace{1cm} (5)

where $A$ and $\phi$ are the harmonic amplitudes and phases at frame $i$, respectively. $F_s = 16$ kHz is the sampling frequency, $t = 0, 1, \ldots, L$, and $L$ is the frame length. $K$ is the number of harmonics that depend on both contF0 and MVF:

$$K^i = \begin{cases} \text{round}\left( \frac{\text{MVF}^i}{\text{contF0}^i} \right) - 1, & \text{voiced frames} \\ 0, & \text{unvoiced frames} \end{cases}$$  \hspace{1cm} (6)

$$A_i^k = 2\sqrt{\text{contF0}^i \cdot H_b^i \left( k \text{contF0}^i \right) \cdot \text{exp}\left( \text{Re}\{C_i^k\} \right)}$$  \hspace{1cm} (7)

where $H_b$ is a complementary low-pass filter for the harmonic part, $C_i$ is complex harmonic log-amplitude obtained by resampling the MGC envelope:

$$C_i^k = c_0^i + 2 \sum_{n=1}^{N} c_n^i \cos\left( n\beta_n^i \right)$$  \hspace{1cm} (8)

Figure 3. Given a speech frame, the sinusoidal model measures the F0 and the MGC representation of the spectral envelope (blue), and the MVF defined as the boundary between harmonic and noisy bands. Frequency bins are intervals between samples in the frequency domain.
\[ \beta_x(w'_0) = \tan^{-1}\left(\frac{(1-x^2) \sin w'_0}{1 + x^2 \cos w'_0 - 2x}\right) \]  \hspace{1cm} (9)

where \(a\) is the all-pass factor that takes 0.42 for \(F_s = 16\ kHz\). The phases are obtained recursively in a minimum phase response between harmonics in adjacent frames:

\[ q'^i_k = 1m\left\{C_k^i\right\} + k\gamma^i \]  \hspace{1cm} (10)

\[ \gamma^i = \gamma^{i-1} + \frac{T}{2}(w'_0 + w'^{i-1}_0) \]  \hspace{1cm} (11)

where \(k\gamma^i\) is a linear-in-frequency term which can be attributed to the underlying excitation, and \(T\) is the frame shift measured in samples (typically it corresponds to a 5\ms interval).

The synthetic noise part \(n(t)\), with this work building on our previous study \([32]\), is firstly filtered by a high-pass filter \(f_h(t)\) with a cutoff frequency equal to the local MVF, and then modulated by its Hilbert envelope \(e(t)\):

\[ s'_n(t) = e'(t) \left[ f'_h(t) + n'(t) \right] \]  \hspace{1cm} (12)

In unvoiced frames, on the other hand, the harmonic part is zero (MVF = 0) and the synthetic frame is formed only by noise. Hence, we can reconstruct the speech signal exactly by summing up the harmonic and noise components.

### 4. Modified Generative Adversarial Networks

Our proposed algorithm is adopted from the StarGAN approach \([33]\), which was proposed for multi-domain image-to-image translation. It differs from the method in \([25]\) by introducing constructive changes in terms of training architecture to be able to adapt the proposed sinusoidal framework. Our aim is to use a single generator \(G\) that can learn mappings between a group of speakers. To achieve this, we train \(G\) to convert the attribute of source \(x \in R^{Q \times T_x}\) speaker domain into target \(y \in R^{Q \times T_y}\) speaker domain conditioned on the target domain label \(c \in \{1, \ldots, N\}\) to generate a new acoustic feature sequence \(\hat{y} = G(x, c)\), where \(N\) is the number of domains, \(Q\) is the feature dimension, and \(T\) is the sequence length. An auxiliary classifier is also introduced as with \([34]\) which allows a real/fake discriminator \(D\) to learn the best decision boundary between the converted and real acoustic features. Hence, our \(D\) produces a probability \(D(\hat{y}, c')\) over both sources and domain labels that \(c'\) (of the input source speech) designed to produce class probabilities \(p(c'|y)\). Figure 4 displays the training process of the suggested approach.

![Figure 4](image-url)
Converted speech can also be clipped (which inevitably changes the spectrum of speech signals), depending on the input gain. This will partially distort the speaker information contained in the signal. To stabilize the training procedure, we applied three preservation losses (adversarial, classification, and reconstruction) in the objective function to alleviate the issues of the over-smoothing caused by statistical averaging. The smaller these losses are, the closer the converted data distribution is to a normal speech distribution.

4.1. Adversarial Loss ($\mathcal{L}_{adv}$)

This loss works to render the converted features indistinguishable from the real target feature:

$$
\mathcal{L}_{adv}(D) = \mathbb{E}_y[\log D(y, c)] + \mathbb{E}_x, c'[\log(1 - D(G(x, c'), c'))]
$$

(13)

$$
\mathcal{L}_{adv}(G) = \mathbb{E}_x, c[\log D(G(x, c), c)]
$$

(14)

where $\mathbb{E}$ means the expected value, $G(x, c)$ generates fake data conditioned on both the source speaker’s data $x$ and the target label $c$, whilst $D$ aims to distinguish between real and fake data. That is, the $G$ seeks to minimize this loss, while the $D$ attempts to maximize it.

4.2. Classification Loss ($\mathcal{L}_{cls}$)

In order to synthesize the acoustic feature that belongs to the target domain, we append an auxiliary classifier $C$ on top of $D$ and impose the $\mathcal{L}_{cls}$ when updating both $D$ and $G$. Thus, we decompose this loss into: a) classification loss of real speech data $\mathcal{L}^\text{real}_{cls}$ used to optimize $D$:

$$
\mathcal{L}^\text{real}_{cls}(C) = \mathbb{E}_{y, c'}[-\log D(c'|y)]
$$

(15)

where the term $D(c'|y)$ represents a probability distribution of real speech data $y$ over domain labels computed by $D$. By minimizing this loss, $D$ tries to classify $y$ to its corresponding domain $c'$. We assume that the input data and domain labels are provided by training examples. Moreover, b) classification loss of fake speech data $\mathcal{L}^\text{fake}_{cls}$ used to optimize $G$:

$$
\mathcal{L}^\text{fake}_{cls}(G) = \mathbb{E}_x, c[-\log D(c|G(x, c))]
$$

(16)

where $D(c|G(x, c))$ represents the probability distribution of fake data $G(x, c)$ over domain labels computed by $D$. Hence, the idea is to minimize $\mathcal{L}^\text{real}_{cls}(C)$ with respect to $C$ and $\mathcal{L}^\text{fake}_{cls}(G)$ with respect to $G$ to construct data that can be classified as the target domain $c$.

4.3. Reconstruction Loss ($\mathcal{L}_{rec}$)

The $\mathcal{L}_{adv}$ and $\mathcal{L}_{cls}$ encourage a converted acoustic feature to become realistic and classifiable, respectively. However, they do not guarantee that the converted feature will preserve the content of the linguistic information while changing only the speaker domain-related information. To alleviate this problem, the $\mathcal{L}_{rec}$ is used:

$$
\mathcal{L}_{rec}(G) = \mathbb{E}_{x, c, c'}[||x - G(G(x, c), c')||_1]
$$

(17)

This $\mathcal{L}_{rec}$ encourages $G$ to find an optimal source and target pair that does not compromise the composition, where $||\cdot||_1$ means L1 norm, $G(x, c)$ is the generated data conditioned on $x$ and the target domain label $c$, and $G(G(x, c), c')$ is reconstructing the original speech $x$ which is conditioned on $G(x, c)$ and the original domain label $c'$. 

4.4. Full Objective

By combining the above three losses, an exemplar can be learned from unparalleled training examples to map an input $x$ to the desired output $y$. More precisely, the objective functions to optimize $G$ and $D$ by minimizing $\mathcal{L}_G$ and $\mathcal{L}_D$ are:

$$
\mathcal{L}_G = \mathcal{L}_{adv} + \lambda_{cls}\mathcal{L}^\text{fake}_{cls} + \lambda_{rec}\mathcal{L}_{rec}
$$

(18)
\[
\mathcal{L}_D = \mathcal{L}_{\text{adv}} + \lambda_{\text{cls}} \mathcal{L}_{\text{cls}}^{\text{real}}
\]
(19)
\[
G^*, D^* = \arg\min_G \max_D (\mathcal{L}_G + \mathcal{L}_D)
\]
(20)

where \(\lambda_{\text{cls}} \geq 0\) and \(\lambda_{\text{rec}} \geq 0\) are regularization parameters of the classification and reconstruction losses, respectively, compared to the adversarial loss. In this study, we use \(\lambda_{\text{cls}} = 1\) and \(\lambda_{\text{rec}} = 10\).

4.5. Conversion Process

First, we feed the source speech into the CSM analyzer to extract the contF0, MVF, and spectral envelope. Then, we build an acoustic feature vector by stacking those parametric features. To have zero mean and unit variance in the training dataset, we normalize the acoustic features over all the speakers. Next, both generator and discriminator are optimized in an iterative way in the training time (number of epochs), where one module is being updated while the model parameters of another are fixed. It can be stated that the contF0 and MVF are converted frame by frame using the statistical distribution (mean–variance transformations) of the source speaker to target one. For example, to convert the contF0:

\[
contF'_{0i} = \frac{\sigma_y}{\sigma_x} (contF_{0i} - \mu_x) + \mu_y
\]
(21)
\[
\mu_{\text{contF0}} = \frac{\sum_{i=1}^{N} w_i \text{contF0}_i}{w_i}
\]
(22)
\[
\sigma_{\text{contF0}} = \frac{\sum_{i=1}^{N} w_i (\text{contF0}_i - \mu_{\text{contF0}})^2}{\sum_{i=1}^{N} w_i}
\]
(23)

where \(contF'_{0i}\) is the log-scaled contF0 after conversion at frame \(i\); \(\mu_x\) and \(\sigma_x\) are the source mean and variance of the contF0, respectively; \(\mu_y\) and \(\sigma_y\) are the target mean and variance of the contF0, respectively; and \(\sum_{i=1}^{N} w_i = 1\) is the weight.

For the spectral envelope, we use low-dimensional representation in a 36-MGC domain to reduce complexity. Thus, the MGC features of the source speaker are converted into those of the target speaker using the trained model described in Section 4. Once the training process is completed, we use the CSM synthesizer to generate speech from the converted features. The converted speech will present both the linguistic contents of the source utterance as well as the speaker attributes (identity, gender, and accent). An algorithm is presented below for summarizing the process of our acoustic modeling.

4.6. Network Architecture

We use a 2D convolutional neural network (CNN) to build the generator and discriminator. The generator network comprises two convolutional layers, six CNN residual blocks [35], two transposed CNN layers, and a 2 × 2 stride for downsampling and upsampling. In contrast, two separate five CNN layers are used for discriminator and classifier networks. Instance normalization [36] is used for the generator as this greatly improves the stability of the training but no normalization is employed for the discriminator. All these layers are followed by a rectified linear unit (ReLU) activation function, and the output layer followed by a sigmoid activation function. Each model is trained using Adam as the optimizer [37] with \(\beta_1 = 0.9\) and \(\beta_2 = 0.999\). The batch size is set to 32.
Algorithm 1 Acoustic modeling for voice conversion

Require: Feature extraction and initialization
1: \( x := \text{source features}; \ c := \text{speaker label}; \ y := \text{target features} \)
2: \( \text{Set parameters } G; \ D \)
3: \( \text{Initialize batch size } m; \ \text{learning rate } \eta; \ \text{weights for losses } \lambda; \ \text{number of total iterations } n \)

Begin: Adversarial training model
1: \( \text{for epoch } \ n_1, \ldots, n_k \) do
2: \( \text{for training examples in } (x, c, y) \) do
3: \( \text{for } i = 1, \ldots, m \) do
4: \( \hat{y} = G(x, c) \)
5: \( D(\hat{y}, c') \)
6: \( \text{end for} \)
7: \( \text{optimize } G \) and \( D \) by minimizing losses \( L_G \) and \( L_D \):
8: \( L_G = L_{adv} + \lambda_{cls} L_{cls}^{fake} + \lambda_{rec} L_{rec} \)
9: \( L_D = L_{adv} + \lambda_{cls} L_{cls}^{real} \)
10: \( G^*, D^* = \arg\min_{G,D} L(G, D) \)
11: \( \text{Update } D \) while fixing \( G \)
12: \( \text{end for} \)

End

Begin: Generation of converted speech
1: \( \text{for training data in } (x, c', y) \) do
2: \( \text{generate } \text{contF0}', \ \text{MVF}', \ \text{MGC}' \)
3: \( \text{synthesis CSM(contF0', MVF', MGC')} \)
4: \( \text{end for} \)

End

5. Evaluation and Discussion

In this section, experimental setup, statistical evaluations, and a perceptual listening test in order to confirm the performance of the proposed framework are described.

5.1. Experimental Setup

Unlike [25, 33], this work is conducted on the CSTR VCTK corpus [38], which contains 46 h of English speech from 109 speakers with various accents. From this database, our proposed system was evaluated on data of 8 speakers (4 males and 4 females). The training examples of each speaker are split into 90% training and 10% test sets. As the conversion setting is non-parallel, we did not use any time alignment procedures for training and each speaker reads a different set of sentences. All the speech signals were resampled to a sampling rate of 16 kHz. The 36-dimensional MGC, 1-dimensional MVF, and 1-dimensional contF0 are extracted from the speech of the speakers, and the corresponding stats (means and standard deviations) computed. The acoustic features were computed using a 25 ms window and 5 ms frameshift. We conduct 8 inter-gender (female-to-male and male-to-female) conversions and 4 intra-gender (male-to-male and female-to-female) conversions.

As there are 8 speakers involved in our experiments, \( c \) is represented as an 8-dimensional one-hot vector and there were 12 different combinations of source and target speakers in total. Thus, there are 960 samples (10 utterances * 8 speakers * 12 tests) that needed to be evaluated in total. An NVidia Titan X GPU is used for training. The state-of-the-art system in this study follows the structure of the StarGAN-VC model [25] that achieved greater naturalness of the converted speech in many-to-many non-parallel VC. Thus, we assess our proposed system with it. For comparative evaluation, we choose to exclude any parallel data in the training and testing data.
5.2. Statistical Evaluations

To show that our model is able to convert speaker characteristics, several performance metrics are used. While the spectral distortion is designed to compute the distance between two power spectra, the root mean square (RMS) log spectral distance (LSD) metric is suggested here to carry out the evaluation as

$$LSD_{RMS} = \sqrt{\frac{1}{N} \sum_{k=1}^{N} [\log P(f_k) - \log \hat{P}(f_k)]^2}$$ (24)

where \(k\) is the frequency bin, \(P(f)\) is the spectral power magnitude of the real speech, while \(\hat{P}(f)\) is the spectral power magnitude of the converted speech, and both are defined at \(N\) frequency points. The optimal value of \(LSD_{RMS}\) is zero, which indicates a matching frequency content. Here, we average the \(LSD_{RMS}\) values for the whole set of tested inter-gender examples (female-to-male and male-to-female) separately as

$$Average\ LSD_{RMS} = \frac{1}{M} \sum_{i=1}^{M} LSD_{RMS_i}$$ (25)

where \(M\) is the total number of examined examples, which is equal to 10 in this study.

We also use a set of example spectrograms for the converted voices, to see whether voice conversion results capture the target speaker. The results are shown in Figure 5, and we found that the proposed framework has lower \(LSD_{RMS}\) values equal to 1.7 dB (female to male) and 2 dB (male to female) that are closer to the target speech spectrogram than the StarGAN model. Consequently, our proposed system introduces a smaller distortion to the sound quality and approaches a correct spectral criterion.

![Figure 5. Example of the spectrogram for source, target, and converted speech. The LSD values are averages for the whole set of tested examples.](image)

In addition, the empirical cumulative distribution function (ECDF) [39] of the phase distortion mean (PDM) [40] is computed and presented in Figure 6. The reason to compute this function is to see whether these conversion methods can be normally distributed and

...
if we can compare it to the natural target speech. PDM is estimated in this experiment at a 5 ms frameshift by (http://covarep.github.io/covarep/ accessed on 24 December 2016):

\[ PDM = \sigma_i(f) = \sqrt{-2 \log \left( \frac{1}{N} \sum_{n \in C} e^{(PD_n(f) - \mu_n(f))} \right)} \]  

where \( C = \{i - \frac{N-1}{2}, \ldots, i + \frac{N-1}{2}\} \), \( N \) is the number of frames, \( PD \) is the phase difference between two consecutive frequency components, and we denote the phase by \( \angle \). The standard deviation in Equation (26) is supposed to represent the noisiness of the voice source, which allows for obtaining a more robust estimate of the source shape in transients. Additionally, conventional systems had issues with modeling the high frequency voiced/voiceless contents, so we wanted to show that our system is better in this sense.

![Figure 6. Empirical cumulative distribution function of PDMs using 4 types of VC pairs evaluated with the natural target speech.](image)
As we wanted to quantify the noisiness in the higher frequency bands only, we zero out the PDD values below the MVF contour. The ECDF $F_n(x)$ is expressed as

$$F_n(x) = \frac{\# \{X_i : X_i \leq x \}}{n} = \frac{1}{n} \sum_{i=1}^{n} I_{X_i \leq x}(X_i)$$

where $X_1, \ldots, X_n$ are the PDM variables, $n$ is the number of experimental observations, $\#A$ represents the number of elements in the sample $X \leq x$, $I$ is the indicator function

$$I(x) = \begin{cases} 1, & x \in A \\ 0, & x \notin A \end{cases}$$

In the positive $x$-axis distribution shown in Figure 6, the proposed system is better reconstructed than that of the StarGAN approach. These outcomes can be supported by the fact that developing a sinusoidal VC system is beneficial and can substantially transform the source speaker into a particular target without parallel corpora. In a negative $x$-axis distribution, as shown in Figure 6, the proposed system still yields better voice conversion performance than others and almost reaches the target distribution for intra-gender pairs (male/male or female/female in Figure 6). Hence, the experimental results objectively validate the success of the sinusoidal VC system, and it offered better performance than state-of-the-art StarGAN.

### 5.3. Qualitative Evaluations

A perceptual listening test was created to assess the performance of our proposed model on a non-parallel many-to-many VC task. We conducted a speaker similarity test to compare the dissimilarity and quality of the converted speech to a natural source speaker. The listeners should give a score for each stimulus, from 0 (highly similar to the source speaker) to 100 (highly similar to the sound of the target speaker). The listeners only have a reference example from both source and target speakers in order to distinguish and know the target speaker. Then, listeners have to evaluate the heard sentences based on the target speaker information (gender, accent, volume, etc.). Different sentences were randomly selected from each conversion pair and given in a randomized order. Thus, 81 utterances were involved in the listening test (3 types x 27 sentences) and randomly presented to the participants. Fifteen participants (seven males and eight females) participated in the experiment. The listening test took roughly 12 min. The audio samples can be found online (https://malradhi.github.io/contSM-VC accessed on 14 June 2021).

Figure 7 shows the scores of the similarity test. The error bars represent the 95% confidence interval. As the results show, the developed and reference systems accomplish the same performance to the target voice. In other words, the proposed model has acceptably transformed the source voice to the target voice in the same-gender and cross-gender cases. It can also be found that both systems have equal performance considering similarity to the target speaker. In other words, there is no statistically significant difference between the proposed and StarGAN systems. The proposed framework validates the effectiveness of the sinusoidal model with continuous fundamental frequency in the conversion pipeline. From this perceptual test, we can conclude that our model is able to produce a voice like that of the target speaker in comparison to the system with more complicated discontinuous F0.
6. Conclusions

This paper proposed a novel alternative framework for advancing the accuracy of non-parallel many-to-many voice conversions. The main idea was to employ a sinusoidal model with continuous parameters to generate converted speech signals with an adversarial training network. The main advantages of the sinusoidal model are the high accuracy of harmonic parameter estimation and increased fidelity in converting the source speaker to the target speaker. From the empirical studies, it was confirmed that the proposed approach has a better ability to convert the source speaker to the target one than the state-of-the-art system. The results of the listening test also revealed the effectiveness of the suggested system for improving the quality of synthetic speech comparable with StarGAN.

As future directions, we will attempt to increase the usability of our sinusoidal approach by using the Griffin–Lim algorithm [41] and $d$-vector technique [42]. The GLA is a phase reconstruction method that involves an iterative procedure for estimating a signal from the short-time Fourier transform (STFT) magnitude. The GLA is based only on the consistency and does not take any prior knowledge about the target signal into account. We also plan to adopt a deep vector or “d-vector” method for voice conversion. A d-vector is a fixed-dimensional representation of a speech utterance that enables more natural voice conversion. We hope that our work enables users to generalize more VC tasks (e.g., multi-emotion and multi-pronunciation).

**Author Contributions:** Conceptualization, M.S.A.-R. and T.G.C.; Formal analysis, M.S.A.-R. and T.G.C.; Investigation, M.S.A.-R.; Methodology, M.S.A.-R. and T.G.C.; Project administration, G.N.; Resources, M.S.A.-R.; Software, M.S.A.-R. and T.G.C.; Supervision, G.N.; Funding acquisition, G.N.; Writing—original draft, M.S.A.-R.; Writing—review and editing, T.G.C. and G.N. All authors have read and agreed to the published version of the manuscript.

**Funding:** The research was partly supported by the European Union’s Horizon 2020 research and innovation programme under grant agreement No. 825619 (AI4EU), and by the National Research Development and Innovation Office of Hungary (FK 124584 and PD 127915).

**Institutional Review Board Statement:** Not applicable.
References

1. Mohammadi, S.H.; Kain, A. An overview of voice conversion systems. Speech Commun. 2017, 8, 65–82. [CrossRef]
2. Kain, A.; Macon, M.W. Spectral voice conversion for text-to-speech. In Proceedings of the IEEE ICASSP, Washington, DC, USA, 15 May 1998; pp. 285–288.
3. Toda, T.; Nakagiri, M.; Shikano, K. Statistical voice conversion techniques for body-conducted unvoiced speech enhancement. IEEE Trans. Audio Speech Lang. Process. 2012, 20, 2505–2517. [CrossRef]
4. Li, A.; Peng, R.; Zheng, C.; Li, X. A Supervised Speech Enhancement Approach with Residual Noise Control for Voice Communication. Appl. Sci. 2020, 10, 2894. [CrossRef]
5. Nakamura, K.; Toda, T.; Saruwatari, H.; Shikano, K. Speaking-aid systems using GMM-based voice conversion for electrolaryngeal speech. Speech Commun. 2012, 54, 134–146. [CrossRef]
6. Das Chakladar, D.; Kumar, P.; Mandal, S.; Roy, P.P.; Iwamura, M.; Kim, B.-G. 3D Avatar Approach for Continuous Sign Movement Using Speech/Text. Appl. Sci. 2021, 11, 3439. [CrossRef]
7. Kobayashi, K.; Toda, T.; Doi, H.; Nakano, T.; Goto, M.; Neubig, G.; Sakti, S.; Nakamura, S. Voice timbre control based on perceived age in singing voice conversion. IEICE Trans. Inf. Syst. 2014, E97-D, 1419–1428. [CrossRef]
8. Stylianou, Y.; Cappe, O.; Moulines, E. Continuous probabilistic transform for voice conversion. IEEE Trans. Speech Audio Process. 1998, 6, 131–142. [CrossRef]
9. Pilkington, N.; Zen, H.; Gales, M. Gaussian process experts for voice conversion. In Proceedings of the Interspeech, Florence, Italy, 27–31 August 2011; pp. 2761–2764.
10. Erro, D.; Moreno, A.; Bonafonte, A. Voice conversion based on weighted frequency warping. IEEE Trans. Audio Speech Lang. Process. 2010, 18, 922–931. [CrossRef]
11. Wu, Z.; Virtanen, T.; Chng, E.S.; Li, H. Exemplar-based sparse representation with residual compensation for voice conversion. IEEE/ACM Trans. Speech Audio Process. 2014, 22, 1506–1521.
12. Nakashika, T.; Takiguchi, T.; Ariki, Y. Voice conversion based on speaker-dependent restricted Boltzmann machines. IEICE Trans. Inf. Syst. 2014, 97, 1403–1410. [CrossRef]
13. Desai, S.; Black, A.; Yegnanarayana, B.; Prahallad, K. Spectral mapping using artificial neural networks for voice conversion. IEEE/ACM Trans. Speech Audio Process. 2010, 18, 954–964. [CrossRef]
14. Al-Radhi, M.S.; Csapo, T.G.; Németh, G. Continuous vocoder applied in deep neural network based voice conversion. Multimedia Tools Appl. 2019, 78, 33549–33572. [CrossRef]
15. Sun, L.; Kang, S.; Li, K.; Meng, H. Voice conversion using deep bidirectional long short-term memory based recurrent neural networks. In Proceedings of the IEEE ICASSP, Brisbane, Australia, 19–24 April 2015; pp. 4869–4873.
16. Kobayashi, K.; Hayashi, T.; Tamamori, A.; Toda, T. Statistical voice conversion with WaveNet based waveform generation. In Proceedings of the Interspeech, Stockholm, Sweden, 20–24 August 2017; pp. 1138–1142.
17. Shah, N.J.; Patil, H.A. A novel approach to remove outliers for parallel voice conversion. Comput. Speech Lang. 2019, 58, 127–152. [CrossRef]
18. Saito, D.; Yamamoto, K.; Minematsu, N.; Hirose, K. One-to-many voice conversion based on tensor representation of speaker space. In Proceedings of the Interspeech, Florence, Italy, 27–31 August 2011; pp. 653–656.
19. Ohtani, Y.; Toda, T.; Saruwatari, H.; Shikano, K. Non-parallel training for many-to-many eigenvoice conversion. In Proceedings of the IEEE ICASSP, Dallas, TX, USA, 14–19 March 2010; pp. 4822–4825.
20. Morise, M.; Yokomori, F.; Ozawa, K. WORLD: A vocoder-based high-quality speech synthesis system for real-time applications. IEICE Trans. Inf. Syst. 2016, 99, 1877–1884. [CrossRef]
21. Espic, F.; Valentini-Botinhao, C.; King, S. Direct Modelling of Magnitude and Phase Spectra for Statistical Parametric Speech Synthesis. In Proceedings of the Interspeech, Stockholm, Sweden, 20–24 August 2017; pp. 1383–1387.
22. Oord, A.V.D.; Dieleman, S.; Zen, H.; Simonyan, K.; Vinyals, O.; Graves, A.; Kalchbrenner, N.; Senior, A.W.; Kavukcuoglu, K. WaveNet: A generative model for raw audio. arXiv 2016, arXiv:2106.16036v2.
23. Saito, Y.; Iijima, Y.; Nishida, K.; Takamichi, S. Non-parallel voice conversion using variational autoencoders conditioned by phonetic posteriorgrams and d-vectors. In Proceedings of the IEEE ICASSP, Calgary, AB, Canada, 15–20 April 2018; pp. 5274–5278.
24. Kaneko, T.; Kameoka, H. CycleGAN-VC: Non-parallel voice conversion using cycle-consistent adversarial networks. In Proceedings of the EUSIPCO, Rome, Italy, 3–7 September 2018; pp. 2114–2118.
25. Kameoka, H.; Kaneko, T.; Tanaka, K.; Hozo, N. StarGAN-VC: Non-parallel many-to-many Voice Conversion Using Star Generative Adversarial Networks. In Proceedings of the IEEE SLT, Athens, Greece, 18–21 December 2018; pp. 266–273.
26. Paul, D.; Pantazis, Y.; Stylianou, Y. Non-parallel Voice Conversion using Weighted Generative Adversarial Networks. In Proceedings of the Interspeech, Graz, Austria, 15–19 September 2019; pp. 659–663.
27. McAulay, R.J.; Quatieri, T.F. Speech analysis/synthesis based on a sinusoidal representation. *IEEE Trans. Acoust. Speech Signal Process.* **1986**, *34*, 744–754. [CrossRef]

28. Garner, P.N.; Cernak, M.; Motlicek, P. A simple continuous pitch estimation algorithm. *IEEE Signal Process. Lett.* **2013**, *20*, 102–105. [CrossRef]

29. Drugman, T.; Stylianou, Y. Maximum Voiced Frequency Estimation: Exploiting Amplitude and Phase Spectra. *IEEE Signal Process. Lett.* **2014**, *21*, 1230–1234. [CrossRef]

30. Masanori, M. CheapTrick, a spectral envelope estimator for high-quality speech synthesis. *Speech Commun.* **2015**, *67*, 1–7.

31. Flanagan, J.L.; Golden, R.M. Phase vocoder. *Bell Syst. Tech. J.* **2009**, *45*, 1493–1509. [CrossRef]

32. Al-Radhi, M.S.; Csapó, T.G.; Németh, G. Time-domain envelope modulating the noise component of excitation in a continuous residual-based vocoder for statistical parametric speech synthesis. In *Proceedings of the Interspeech*, Stockholm, Sweden, 20–24 August 2017; pp. 434–438.

33. Choi, Y.; Choi, M.; Kim, M.; Ha, J.W.; Kim, S.; Choo, J. StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE Conference on CVPR*, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8789–8797.

34. Odena, A.; Olah, C.; Shlens, J. Conditional Image Synthesis with Auxiliary Classifier GANs. In *Proceedings of the ICML*, Sydney, Australia, 6–11 August 2017; pp. 2642–2651.

35. Ulyanov, D.; Vedaldi, A.; Lempitsky, V. Instance normalization: The missing ingredient for fast stylization. *arxiv 2016*, arXiv:1607.08022.

36. Odena, A.; Olah, C.; Shlens, J. Conditional Image Synthesis with Auxiliary Classifier GANs. In *Proceedings of the ICML*, Sydney, Australia, 6–11 August 2017; pp. 2642–2651. [CrossRef]

37. Waterman, M.S.; Whiteman, D.E. Estimation of probability densities by empirical density functions. *Int. J. Math. Educ. Sci. Technol.* **1978**, *9*, 127–137. [CrossRef]

38. Christophe, V.; Junichi, Y.; Kirsten, M.D. CSTR VCTK Corpus: English Multi-Speaker Corpus for CSTR Voice Cloning Toolkit; University of Edinburgh, The Centre for Speech Technology Research (CSTR): Edinburgh, UK, 2017.

39. Variani, E.; Lei, X.; McDermott, E.; Moreno, I.; Gonzalez-Dominguez, J. Deep neural networks for small foot-print text-dependent speaker verification. In *Proceedings of the IEEE ICASSP*, Florence, Italy, 4–9 May 2014; pp. 4052–4056.