LPCSE: Neural Speech Enhancement through Linear Predictive Coding

Yang Liu*, Na Tang†, Xiaoli Chu‡, Yang Yang§, and Jun Wang¶

*†The University of Sheffield, UK
†University of Chinese Academy of Sciences, China
‡Terminus Group, China
¶University College London, UK

Email: {†yangliu,‡x.chu}@sheffield.ac.uk, †tangna@mail.sim.ac.cn,
§dr.yangyang@terminusgroup.com, ¶jun.wang@cs.ucl.ac.uk

Abstract—The increasingly stringent requirement on quality-of-experience in 5G/B5G communication systems has led to the emerging neural speech enhancement techniques, which however have been developed in isolation from the existing expert-rule based models of speech pronunciation and distortion, such as the classic Linear Predictive Coding (LPC) speech model because it is difficult to integrate the models with auto-differentiable machine learning frameworks. In this paper, to improve the efficiency of neural speech enhancement, we introduce an LPC-based speech enhancement (LPCSE) architecture, which leverages the strong inductive biases in the LPC speech model in conjunction with the expressive power of neural networks. Differentiable end-to-end learning is achieved in LPCSE via two novel blocks: a block that utilizes the expert rules to reduce the computational overhead when integrating the LPC speech model into neural networks, and a block that ensures the stability of the model and avoids exploding gradients in end-to-end training by mapping the Linear prediction coefficients to the filter poles. The experimental results show that LPCSE successfully restores the formants of the speeches distorted by transmission loss, and outperforms two existing neural speech enhancement methods of comparable neural network sizes in terms of the Perceptual evaluation of speech quality (PESQ) and Short-Time Objective Intelligibility (STOI) on the LJ Speech corpus.

Index Terms—speech enhancement, linear predictive coding, neural networks, formant distortion

I. INTRODUCTION

With the development of 5G/B5G communication systems and the rise of artificial intelligence [1], high-quality speech communications are required in various applications, such as Automatic Speech Recognition (ASR), Augmented Reality (AR), Unmanned Aerial Vehicle (UAV) communications [2], and semantic communications [6]. However, the speech will inevitably be interfered with by the process of transmission, such as background noise, sound transmission loss (STL) [7], and multipath fading [4]. Particularly, a speech usually suffers from an obvious frequency-dependent STL when transmitting through objects, such as water [5], face masks [8], bone conduction [17], and walls [7]. STL will distort the speech formants which are the most sonorant components in a syllable, and will detrimentally reduce intelligibility.

This work was supported in part by the European Union’s Horizon 2020 research and innovation programme under grant agreement No. 898893.

Speech enhancement (SE) aims at improving the quality and intelligibility of target speech and suppressing unwanted distortions [9]–[12]. Classic expert-rule based methods tried to enhance the formant-distorted speech, such as Linear Predictive Coding (LPC) [21], and Wiener filtering [25], but their expressive powers are typically restricted. With the development of machine learning (ML), many neural SE networks have been proposed for denoising or dereverberation [13], [18], [28]. They exploit the well-structured spectrums or the periodic tones as the input features for SE, but the features are usually damaged due to the frequency-dependent STL, which brings challenges to existing neural SE methods. Furthermore, most purely data-driven ML approaches are unable to extract interpretable knowledge from data and may be physically inconsistent or implausible [16]. To address this challenge, some initial works are focusing on combining expert-rule based models with neural networks, such as the grey box system simulation [3], and the LPC-based speech vocoder [19].

Despite achieving impressive results by neural SE, three challenges remain under-explored. Firstly, although many expert-rule based technologies work on various types of distortions, most neural network-based efforts still focus on noise suppression or dereverberation, while how to use neural networks against formant distortion has not been well studied. Secondly, it is not straightforward to design a simple and compact SE network to identify and estimate the effects of STL on speech without assuming any prior knowledge of the transmission channel. Last but not least, most neural networks are often inefficient since they do not utilize the existing knowledge of how speech is generated or distorted, but how to use the strong inductive biases in the rules of speech pronunciation and distortion without losing the expressive power of neural networks remains an open problem.

In this paper, we propose an LPC-based single-channel SE network architecture (LPCSE), which integrates an interpretable LPC speech model into auto-differentiable neural networks and combines the strong inductive biases in expert rules with the benefits of neural networks. The experiments on the LJ Speech corpus demonstrate that LPCSE successfully restores the formants distorted by the STL without the need
for large-scale, complex networks or models, and obtains significant gains in terms of the Perceptual evaluation of speech quality (PESQ) [22] and Short-Time Objective Intelligibility (STOI) [26] as compared with two existing neural SE methods of comparable network sizes. The contributions of this paper are summarized as follows:

- We introduce LPCSE, an expert-rule inspired network architecture, to perform end-to-end SE for formant-distorted speeches. LPCSE enables utilizing the strong inductive biases of the LPC speech model while retaining the expressive power of SE neural networks and the benefits of end-to-end learning.
- We propose two new blocks to overcome the large computational overhead when integrating the LPC speech model into neural networks and ensure the stability of the model in end-to-end learning, respectively. Based on the blocks, neural networks are used to mitigate the impact of STL on speeches.
- Our ablation and comparison experiments on the LJ Speech corpus verify that the combination of the LPC speech model and neural networks in LPCSE can utilize the existing knowledge of speech generation and STL distortion as well as the expressive power of neural networks, which provide a larger gain in terms of speech quality and intelligibility than each of them working alone. The LPCSE outperforms two existing neural SE methods of comparable network sizes in terms of PESQ and STOI.

II. LPC-BASED NEURAL SPEECH ENHANCEMENT

A. Background - the Expert Rules of LPC

Speech is the production of our phonatory systems, including lungs, larynx, vocal tract, etc., which can be well modeled by an LPC speech model [20]. The vocal tract response can be interpreted as an LPC filter, i.e., a time-varying all-pole linear filter with the technique of LPC. As shown in Fig. 1, speech $s(k)$ is produced by passing an excitation signal $n(k)$ through the LPC filter, and can be written as follows:

$$ s(k) = \sum_{p=1}^{P} a_p s(k-p) + n(k), $$

where $k$ denotes the time index, $P$ is the order of the filter, $a_p$, $p = 1, ..., P$, represents the LP coefficients, which hinge on the vocal tract and remain unchanged for short segments of speech, e.g., a few milliseconds. The excitation signal $n(k)$ is related to the pronunciations of words and can be either a train of periodic impulses for voiced sounds, such as nasals and vowels, or a random noise source for unvoiced sounds. By analyzing speech $s(k)$ and its past samples with LPC, the LP coefficients $a_p$ and excitation $n(k)$ can be determined from short segments of speech by minimizing the sum of squared differences between the actual and predicted speech samples.

The speech observed by a receiver, denoted by $x(k)$, is the convolution of the speech $s(k)$ and the channel impulse response $h$ of the environment plus additive noise $b(k)$, i.e.,

$$ x(k) = s(k) * h + b(k), $$

where $*$ represents linear convolution. For an environment with fixed distortion, such as speech transmission through a wall, it is usually assumed that $h$ is time-invariant (or short-time time-invariant) and can be represented as a finite impulse response (FIR) filter [7]. To find out the relationships between formant-distorted speeches and their clean pairs, it is assumed that the excitation signal $n(k)$ remains unchanged after the distortion [17], and the environment only affects the vocal tract response, i.e., the all-pole linear filter. Note that the limited order $P$ of the LPC filter makes it difficult to describe the complex distortion of environments, while the performance of speech enhancement is considerably dependent on the accuracy of the filter estimation. To address this challenge, we design a Mel-spectrogram enhancement network to improve speech quality.

B. Problem Formulation

Our task is to recover $s(k)$, given only the observed speech $x(k)$ without assuming any priori knowledge regarding the channel impulse response $h$. Let $\mathcal{E}$ denote the index set of $|\mathcal{E}|$ possible scenarios. Let $X^e$ be the observed speech, which is the same speech received in scenario $e \in \mathcal{E}$. The speech of $X^{e=1}$ denotes the scenario without suffering from distortion and is used as ground truth data. For $j = 1, \ldots, m$, let $X^e_j$ be the $j$-th sample of $X^e$ received in $e$. $W^e \in \mathbb{R}^{m \times m}$ defines a directed acyclic graph (DAG) on $m$ samples. Specifically, $W^e = \{w^e_1, \ldots, w^e_j, \ldots, w^e_m\}$ is the LP-coefficient matrix of $X^e$ and $w^e_j$ is the LP coefficient of $X^e_j$. There exists a vector of excitation signals $Z = [z_1, z_2, \ldots, z_m]$ that is independent of the scenario $e$ for all $e \in \mathcal{E}$. Hence, $X^e_j$ can be represented as:

$$ X^e_j = w^e_j X^e + z_j, j = 1, \ldots, m. $$

This is similar to the formulation for the structural equation model (SEM) in DAG learning problems [24]. But the weighted adjacency matrix $W$ in SEM is usually assumed to be fixed across all $e \in \mathcal{E}$, and the $W$ in the LPC speech model is dependent on the scenario $e$ and the excitation signals $Z$. Furthermore, the $Z$ in SEM is treated as unknown random noise, but the excitation signals $Z$ can be decoded from the $X^e$ through the Levinson-Durbin recursion [20].

For brevity, we abbreviate $e = 1$ in the symbols of clean speech, and in the rest of this paper, $e \in \mathcal{E}$ and $e \neq 1$. 

![Diagram of speech production and distortion](image_url)
Specifically, the clean speech \( X^c = \mathbf{1} \) and its DAG \( W^c = \mathbf{1} \) are denoted by \( X \) and \( W \), respectively. Their distorted pairs are defined as \( X^e \) and \( W^e \). Hence, based on Eq.(3), the clean speech can be represented as \( X = W^T X + Z \), which can be reformulated as a matrix multiplication as follows:

\[
X = V Z, \quad V = (I - W^T)^{-1} \in \mathbb{R}^{(ML+1) \times (ML+1)},
\]

where \( M \) is a given number of adjacent samples that share the same LP coefficients. The excitation signal \( Z \) is obtained by the Levinson-Durbin recursion and the clean speech and its distorted pairs have the same \( Z \). We refer to \( M \) samples as a slot, and a frame has \( L \) slots. Increasing \( L \) can reduce the number of waveform phase discontinuities and improve speech quality, but will increase computation and memory requirements. To enhance distorted speech \( X^e \), a straightforward strategy is to use a network \( F \) for estimating the matrix \( V \) that minimizes the least square loss,

\[
\min_{\theta} ||V Z - F(X^c; \theta) Z||^2_2,
\]

where \( \theta \) is the parameter set of the network \( F \). However, we find that although most of the speech pitch and formants are recovered by the matrix \( V \) estimated by \( F \), additional noises are introduced to the speech due to the limited order of the LPC speech model and the loss in the LP-coefficient matrix estimation. To address this challenge, another network \( G \) is proposed to remove the noise and further enhance the speech in the time-frequency (TF) domain as follows:

\[
\min_{\varphi} \left\| X_M - G(X^c_M; \varphi) \right\|^2_2
\]

subject to \( \theta^* = \arg\min_{\theta} \left\| X - F(X^c; \theta) Z \right\|^2_2 \)

\[
X_M = M(X), \quad X^e_M = M(F(X^c; \theta^*) Z; X^c),
\]

where \( X_M \) denotes the Mel-spectrograms of the clean speech \( X \), and \( X^e_M \) denotes the Mel-spectrogram of the enhanced speech \( F(X^c; \theta) Z \) concatenated with the distorted speech \( X^e \). \( M \) represents the transformation from the audio waveforms to Mel-spectrogram. \( \varphi \) denotes the parameters of \( G \). The Mel-spectrogram transformation is differentiable and can be implemented by the standard modules in ML frameworks, such as the \texttt{torchudio} module in \texttt{Pytorch} [27].

### III. NETWORK ARCHITECTURE

LPCSE solves the SE problem with a cross-domain architecture, which includes four key components: a data preprocessing step, a time-domain network \( F \) for LP coefficients estimation and waveform generation, a TF domain network \( G \) for Mel-spectrogram enhancement, and a vocoder. In the data preprocessing step, i.e., box (i) in Fig. 2, a distorted speech is decoded into the LP coefficients and residuals with an LPC order of \( P \) and a step size of \( M \) samples. The network \( F \), i.e., box (ii) in Fig. 2, concatenates \( M \times L \) LP coefficients and their residuals as input. In this module, a stack of convolutional and recurrent networks with a \texttt{Tanh} layer is used to extract the features from the concatenated LP coefficients and residuals, and estimate the poles of the LPC filter. Two dedicated blocks, i.e. \texttt{LP2Wav} and \texttt{Poles2LP}, are proposed to transfer the poles to LP coefficients and ensure the stability of the LPC speech model. In the network \( G \), i.e., box (iii) in Fig. 2, the waveform from the \texttt{LP2Wav} block is upsampled to 22 kHz by interpolating to match the sampling rate of the Mel-spectrogram vocoder and its clean speech pairs. In the \texttt{Wave2Mel} block, the upsampled waveform is concatenated with the distorted speech, and transformed into two 80-channel log-Mel spectrograms corresponding to the waveforms using the 2048-point Short-Time Fourier Transform (STFT) with a \texttt{Hamming} window of 50 ms frame length and 12.5 ms frameshift. Finally, the Mel-spectrogram is further enhanced via a convolutional-recurrent network and a \texttt{Post-net} [23] to denoise and improve speech quality. In box (iv) in Fig. 2, the enhanced Mel-spectrograms are converted into an enhanced waveform through \texttt{WaveRNN} [15]. Such an architecture takes advantage of the well-structured Mel-spectrograms and the time-domain knowledge from the LPC speech model. \texttt{LP2Wav} and \texttt{Poles2LP} will be introduced in the next subsections.

#### A. The \texttt{LP2Wav} Block

To achieve end-to-end learning, the LPC speech model will be integrated into the network \( F \) as a whole. Since \( V \in \mathbb{R}^{(ML+1) \times (ML+1)} \) is a dense matrix, directly predicting \( V \) from \( X^e \) is a challenge, owing mainly to the high computational overhead using standard networks to approximate \( V \) when \( V \)‘s size is large. In this subsection, we will address this challenge in two steps: sparse matrix transformation and compression. In the first step, we use an inverse operation to convert the dense matrix \( V \) into a sparse matrix. Specifically, we find that the inverse of \( V \) is a sparse matrix based on the definition of the LPC speech model. Let \( U = I - W^T = V^{-1} \) be the inverse of \( V \), which is always invertible. The derivative of \( V \) can be obtained from \( \bar{V} = (UV)' = UV' + U'V = 0 \) and \( V' = -UV' \), where \( U' = -(W^T)' \) and \( (\cdot)' \) is the derivative operator. \( W \in \mathbb{R}^{(ML+1) \times (ML+1)} \) is a strictly upper triangular sparse matrix, which has 0 along its diagonal as well as the
lower portion. Based on the derivative of $W^T$, the gradient of $V$ is calculated by the chain rule, and the parameters of network $F$ and $\mathcal{G}$ are updated by the backpropagation algorithm. In the second step, we compress the sparse matrix $W$ obtained from the first step into a smaller matrix to remove redundant information and reduce computation. Specifically, $W$ is mapped to a smaller LP-coefficient matrix $A'$. Each column of $A'$ contains the LP coefficients for a speech sample. Then, $A'$ is downsampled to $A \in \mathbb{R}^{P \times L}$ without losing information because the LPC speech model assumes that the samples in a short frame have the same LP coefficients. The downsampling factor is set to $M$. We use the proposed networks to estimate the LP coefficients $A$ of the distorted speech, instead of directly predicting $V$, based on which much performance improvement is achieved. For example, $M$ and $P$ are set empirically to 46 and 11 in our experiments, i.e., $(ML + 1)^2 \gg PL$, so the network size and computational overhead are significantly reduced.

To achieve the transformation from the estimated $A$ to $V$, we design a differentiable block, called $LP2Wave$, as shown in Alg.1, where batch is the batch size. The first step in the algorithm represents the reverse process of the compression from $A'$ to $A$. The second step denotes the mapping from the matrix $A'$ to the sparse matrix $W$, which is achieved by using index and in-place operations without making a copy. In the third step, $I - W^T$ is always invertible and its inverse can be easily calculated by using forward substitution because the index and in-place operations ensure that $I - W^T$ is an upper triangular sparse matrix with 1 along its diagonal. $I$ is the identity tensor with the same shape as $W$. To better understand the $LP2Wave$ block, an example illustrates the compression operation for the sparse matrix $W$, as shown in Fig. 3. For 12 speech samples and $P = 3$, $M$ and $L$ are set to 4 and 3, respectively. $W \in \mathbb{R}^{13 \times 13}$ is downsampled to $A \in \mathbb{R}^{3 \times 3}$, without losing information.

Algorithm 1 The process of the $LP2Wave$ block

**Input:** The estimated LP-coefficient matrix $A(batch, L, P)$.

**Output:** Updated $V(batch, ML + 1, ML + 1)$.

1. Obtain $A'(batch, ML, P)$ through the interpolation and concatenation to $A(batch, L, P)$.
2. Update $W(batch, ML + 1, ML + 1)$ by using the index and in-place operations to $A'(batch, ML, P)$.
3. $V(batch, ML + 1, ML + 1) \leftarrow (I - W(batch, ML + 1, ML + 1)^T)^{-1}$.

B. The Poles2LP Block

When using a network to estimate the LP coefficients for an LPC speech model, the most basic requirement is that the coefficients should ensure the stability of LPC. Lacking stability will introduce infinite values to $V$ and cause exploding gradients in the inverse operation of $I - W^T$. In this subsection, we use standard operations in neural networks to guarantee stability. Based on Eq.(1), the system transfer function of the LPC speech model is an all-pole linear filter, and can be rewritten as follows by using the $z$-transform:

$$l(z) = \frac{Kz^P}{z^P - a_1z^{P-1} - a_2z^{P-2} - \ldots - a_P},$$  (7)

where $K$ is the system gain, which is set to 1 in our experiments. $\{a_1, a_2, \ldots, a_P\}$ are the real LP coefficients. To guarantee stability, $l(z)$’s poles must lie inside or on the unit circle on the complex plane. A straightforward idea is to use a supervised learning model to estimate $\{r_1, r_2, \ldots, r_P\}$ with the constraint of $|r_i| \leq 1$ for $i = 1, 2, \ldots, P$ which will guarantee the stability, instead of directly estimating $\{a_1, a_2, \ldots, a_P\}$. However, how to enable the transformation from the poles $r_i$ to the coefficients $a_i$ in auto-differentiable neural networks remains an open problem.

To bridge the gap, we propose a differentiable block for the transformation in the network $F$, called $Poles2LP$, which maps the poles of $l(z)$ to the LP coefficients in $A$ using only standard operators and blocks in ML, as shown in Fig. 4. Firstly, $l(z)$ can be rewritten as:

$$l(z) = \frac{Kz^P}{(z - r_1)(z - r_2) \ldots (z - r_P)},$$  (8)

where $\{r_1, r_2, \ldots, r_P\}$ denote the poles of $l(z)$, and their real and imaginary parts are constructed by the output of the $Tanh$ layer in the network $F$. Secondly, since polynomial multiplication is equivalent to convolution, the LP coefficients are calculated through the iterative convolution between the vector $[1, -r_i]$ and convolution results, $i = 1, 2, \ldots, P$. A tensor of $\text{ones}(batch, L, 1)$ is concatenated with $-r_i\text{batch}, L, 1)$ in the last dimension. Thirdly, a complex convolution block [13] calculates the iterative convolution between the concatenated tensor and previous convolution results. Finally, after the iteration for the $P$ poles, the real part of the final convolution results is the downsampled LP-coefficient matrix $A(batch, L, P)$, which contains $\{a_1, a_2, \ldots, a_P\}$ in its columns. A flowchart is shown in Fig. 4 to describe the process of the $Poles2LP$ block, where $\Re$ is the real part symbol, and $\text{tmp}$ is the convolution result. $P$ is a small value, e.g., 11 in our experiment. complex_conv denotes the complex convolution.

C. Loss Functions

Our goal is to solve the following optimization problem:

$$\min_{\theta, \phi} L = L_{Mel} + \mu L_{Wav} + \lambda L_{LP},$$  (9)

Fig. 3: The expert rules reduce the computational overhead when integrating the LPC speech model into neural networks.
where $\mu$ and $\lambda$ are hyperparameters chosen for balance. $L_{Mel}$ is a Mel reconstruction loss, $L_{Wave}$ is a waveform reconstruction loss, and $L_{LP}$ is an LP-coefficient loss.

**Mel reconstruction loss.** We refer to the output of the full connect layer in $G$ as the coarse Mel-spectrogram, denoted by $\hat{X}_M^c$. We refer to the sum of the coarse Mel-spectrogram and the output of the Post-net as the fine-grained Mel-spectrogram, denoted by $\hat{X}_M^{c\prime}$, i.e., $\hat{X}_M^{c\prime} = g(M(F(X^f; \theta)Z; X^c); \phi)$. We use the sum of the mean squared error (MSE) between the target Mel-spectrogram $X_M$ and the coarse and fine-grained Mel-spectrograms as the Mel reconstruction loss, i.e., $L_{Mel} = \mathbb{E} \left\| \hat{X}_M^c - X_M \right\|_2^2 + \mathbb{E} \left\| \hat{X}_M^{c\prime} - X_M \right\|_2^2$.

**Waveform reconstruction loss.** We will refer to the upsampled waveform in the network $G$ as the coarse-grained speech, denoted by $\hat{X}_c^c$, i.e., $\hat{X}_c^c = F(\hat{X}_c^c; \theta)Z$. We use the MSE between the target $X$ and upsampled waveform $\hat{X}_c^c$ as the waveform reconstruction loss, i.e., $L_{Wave} = \mathbb{E} \left\| \hat{X}_c^c - X \right\|_2^2$.

**LP-coefficient loss.** We use an LP-coefficient loss to the loss function, denoted by $L_{LP}$ in Eq.(9), which represents the expectation of the complex-valued $L_1$ norm between the estimated LP coefficients and the target LP coefficients, i.e., $L_{LP} = \mathbb{E} \left\| \hat{A}^c - A \right\|_1^2$, where $\hat{A}^c$ and $A$ are the estimated and target LP coefficients, respectively. $\hat{A}^c$ is the input of the $\text{Poles2LP}$ block. $A$ is obtained from the clean speech pairs of $X^c$ in the training dataset. Note that $\hat{A}^c$ is a complex tensor, and $A$ is a real tensor. During training, the imaginary part of $\hat{A}^c$ will approach 0, and the real part will be close to $A$.

### IV. Experiments

#### A. Datasets

In our experiments, the LJ Speech corpus [14] is used as the clean dataset to evaluate the proposed LPCSE. LJ Speech is a public data set and contains 13,100 audio clips of a single speaker reading passages with a sampling rate of 22 kHz. To construct distorted-clean speech pairs, we let the clean speech pass through an FIR filter, which is used to simulate the channel impulse response $h$ of a 5 cm thick concrete wall. The FIR filter is designed according to the Sharpe’s equation [7], which is widely used to estimate the STL of walls in the noise control of buildings. Then, we mix the filtered speech with pink noise at 3 Signal-to-noise ratio (SNR) levels from -3 to 3 dB. The distorted-clean speech pairs are divided into training and test sets by 9:1. We use the PESQ and STOI as the evaluation metrics, which measure the quality and intelligibility of speech, respectively. The hyperparameters $\mu \in [0, \infty]$ and $\lambda \in [0, \infty]$ are set to $\mu = 1$, $\lambda = 0.3$. We train our model on 2 NVIDIA 2080Ti GPUs with a batch size of 16 frames on each GPU. The frame and slot length are set to $L = 120$, and $M = 46$. We use the ADAM optimizer with a fixed learning rate of 0.001.

#### B. Ablation Study

An ablation study is conducted to identify the essential components of the proposed architecture. The ablation results are shown in Fig. 5.

**LPCSE and Distorted:** among these methods, LPCSE represents our full architecture, which includes the proposed network $F$ and $G$. Distorted denotes the formant distorted speech without enhancement. A significant gain of 0.25-0.38 and 0.07-0.11 from LPCSE is observed in median PESQ and STOI when compared with Distorted. The results show that the proposed network can improve speech quality and intelligibility in different scenarios. LPCSE significantly outperforms other methods in the ablation study.

**LPCSE-w/o-$G$:** LPCSE-w/o-$G$ shows the performance of only using the network $F$ and removing the network $G$ from the LPCSE architecture, except for the $LP2Wav$ and upsampling layer in the network $G$. The network $F$ estimates the clean LP coefficients based on the excitation signal and distorted LP coefficients, and the LP coefficients are fed into the $LP2Wav$ block. The output of LPCSE-w/o-$G$ is the upsampled waveform from the output of the $LP2Wav$ block, instead of the Mel-spectrogram from the network $G$. The results in Fig. 5 show that, compared with the LPCSE architecture, LPCSE-w/o-$G$ only provides a small gain of 0.02-0.14 and 0-0.03 on PESQ and STOI, respectively. This is expected because LPCSE-w/o-$G$ only uses a simple traditional method, i.e., the LPC speech model, to synthesize speech, which is difficult to mitigate the noise in generated speech. The aim of LPCSE-w/o-$G$, including the network $F$, $LP2Wav$ block and upsampling layer, is to recover the distorted high-frequency components and provide coarse Mel-spectrograms, based on which the rest of the network $G$ will generate fine-grain Mel-spectrograms.

**LPCSE-w/o-$F$ and LPCSE-w-$F’$:** LPCSE-w/o-$F$ and LPCSE-w-$F’$ are the two settings that remove the network $F$ from the LPCSE architecture, while the network $G$ is preserved. The difference is that the LPCSE-w/o-$F$ uses the same network $G$ as LPCSE, and LPCSE-w-$F’$ uses an enhanced network $G’$, which has twice the size of $G$. We see that LPCSE-w/o-$F$ could provide a gain of 0.1-0.16 and 0.01-0.05 on median PESQ and STOI, respectively. LPCSE-w-$F’$ has a larger capacity for speech enhancement than LPCSE-w/o-$F$, but compared with LPCSE-w/o-$F$, it only introduces a small increase of 0.01-0.08 and 0.02-0.04 on the median PESQ and STOI, respectively. The gains of both settings drop significantly as the distortion and noise become strong, i.e.,
from -3 dB to 3 dB, because the network $G$ is insufficient to enhance the speech alone. Comparison between LPCSE-w/o-$F$ and LPCSE shows that our architecture could provide 0.15-0.22 and 0.05-0.08 gain on median PESQ and STOI in different scenarios. This proves that, compared with the architectures which only use the network $G$, LPCSE has a more stable performance as the distortion and noise increase.

In summary, the results of the ablation study suggest that both the network $F$ and $G$ are crucial to the LPCSE architecture. Jointly integrating the networks in our design significantly outperforms the methods where any of them is applied separately since the combination of them is a more effective way to improve speech quality and intelligibility than only using the network $F$ or $G$.

C. System Comparison

We compare LPCSE with two existing neural SE networks, EHNET [29] and PHASEN [28], in terms of PESQ and STOI, and the number of trainable parameters (Params), as shown in Table I. To compare the complexity of the methods, we also measure the forward and backward (F/B) GPU time for an input speech of 7 seconds. EHNET is a convolutional-recurrent method worked in the TF domain, which only uses the Mel-spectrograms of the distorted speech as input. PHASEN is a cross-domain method that captures both amplitude and phase-related information of distorted speeches. The results show that LPCSE significantly outperforms ENNET and PHASEN in terms of PESQ and STOI because the LPC speech model uses the mathematical model with stability constants that make the search space of the parameters tractable. However, in exchange, LPCSE requires a slightly larger computation of 102 and 164 ms GPU time in the F/B propagation, respectively, and the computation comes mainly from the inverse operation in the $LP2Wav$ block because we use an inverse operation for a large matrix to replace the loop in the LPC model and improve computational efficiency on GPUs. The loop can be easily processed on CPUs as in traditional LPC-based algorithms but makes it difficult to be computed in GPU-based auto-gradient ML frameworks. We will reduce the F/B GPU time through program optimization in future works, such as using C/C++ functions to process the loop and avoiding the inverse operation for the large matrix.

![Fig. 5: PESQ and STOI comparisons of the ablation study.](image)

![Fig. 6: The comparison of Mel-spectrograms, poles and formants for a speech from the test dataset.](image)

![D. Visualizations](image)

To have a better understanding of the LPC speech model in LPCSE, we do a case study by visualizing the enhanced speech. Fig.6(a) shows the Mel-spectrograms, pitch, and formants of a formant distorted speech, and its LPCSE-enhanced, and clean pairs from the test dataset. The formants are calculated by Burg’s algorithm for linear prediction coefficients [30]. The pitches are computed by an acoustic periodicity detection algorithm based on an accurate autocorrelation method [31]. Compared with the clean speech, the high-frequency components of the formant distorted speech are damped by noise, and the pitch is less affected due to the low-frequency components maintained. After the SE through LPCSE, the high-frequency components are recovered in the enhanced speech which restores the formants. The poles of the LPC speech model in a slot of the speeches are shown in Fig.6(b).
Compared with the distorted speech, the poles of the enhanced speech are more close to that of the clean speech, which indicates that the LP coefficients of the enhanced speech are corrected by LPCSE. To intuitively show the performance of formant recovery, the first formant differences between the distorted/enhanced and clean speech are shown as the red and blue blocks in Fig.6(b), respectively. From the formant differences between distorted and clean speech, we can see that the first formant difference has an obvious peak of about -200 Hz. From the formant difference between enhanced and clean speech, we can see that the differences between the formants become small, and the mean frequency differences are close to 0, which indicates most of the formants are restored by LPCSE, and the formants of the enhanced speech are close to that of the clean speech. Note that, many speech recognition algorithms use the formants as the features that recognize the speech contents, and the formants restored by LPCSE will improve the recognition rate of the algorithms.

V. CONCLUSION

In this article, we have investigated an LPC-based SE architecture, named LPCSE, which aims to enhance the speeches distorted by the frequency-dependent STL through the combination of LPC and neural networks. Our experiment results have shown that the proposed LPCSE can correct and recover the formants of the distorted speeches. According to our ablation study, the LPC speech model embedded in LPCSE plays a key role in improving speech quality and intelligibility, thereby demonstrating the effectiveness of the LPC in neural SE. In the comparison with two existing neural SE methods of comparable network sizes, LPCSE has larger improvements in terms of PESQ and STOI, which indicates that LPCSE opens a new path toward simpler and better SE systems with the help of expert rules.

REFERENCES

[1] Y. Yang et al., “6G Network AI Architecture for Everyone-Centric Customized Services,” IEEE Network, early access, July 2022.
[2] N. Tang, H. Tang, B. Li, and X. Yuan, “Cognitive NOMA for UAV-enabled secure communications: Joint 3D trajectory design and power allocation,” IEEE Access, vol. 8, pp. 159 963–159 978, 2020
[3] Z. Lin et al., “DRESIA: Deep reinforcement learning-enabled gray box approach for large-scale dynamic cyber-twin system simulation,” IEEE Open J. Comput. Soc., vol. 2, pp. 321-333, 2021.
[4] Y. Liu, W. Zhang, Y. Yang, W. Fang, F. Qin, and X. Dai, “PAMF: Phase-based acoustic motion tracking in multipath fading environments,” Proc. of IEEE INFOCOM, pp. 2386–2394, 2019.
[5] J. Luo, Y. Chen, M. Wu, and Y. Yang, “A survey of routing protocols for underwater wireless sensor networks,” IEEE Commun. Surv. Tutor., vol. 23, no. 1, pp. 137-160, 2021.
[6] Z. Weng, Z. Qin, and G. Y. Li, “Semantic Communications for Speech Recognition,” Proc. of IEEE GLOBECOM, pp. 1-6, 2021.
[7] D. A. Bies, C. Hansen, and C. Howard, “Engineering Noise Control, 5th ed.” Productivity Press, New York, 2017.
[8] R. M. Corey, U. Jones, and A. C. Singer, “Acoustic effects of medical, cloth, and transparent face masks on speech signals,” J. Acoust. Soc. Am., vol. 148, no. 4, ASA, pp. 2371-2375, 2020.
[9] S. E. Eskimez, T. Yoshioka, H. Wang, X. Wang, Z. Chen and X. Huang, “Personalized speech enhancement: new models and Comprehensive evaluation,” Proc. of IEEE ICASSP, pp. 356-360, 2022.
[10] Li, Andong, et al., “Glance and gaze: A collaborative learning framework for single-channel speech enhancement,” Appl. Acoust. vol. 187 no. 108499, 2022.
[11] Michelsanti, Daniel, et al., “An overview of deep-learning-based audiovisual speech enhancement and separation,” IEEE/ACM Trans. Audio, Speech, Language Process. vol. 29 pp. 1368-1390, 2021.
[12] Hao, Xiang, et al., “Fullsubnet: A full-band and sub-band fusion model for real-time single-channel speech enhancement,” Proc. of IEEE ICASSP, pp. 6633-6637, 2021.
[13] Yanxin Hu, et al., “DCCRN: deep complex convolution recurrent network for phase-aware speech enhancement,” Proc. of Interspeech, pp. 2472-2476, 2020.
[14] “The LI Speech Dataset,” Keithito.com. [Online]. Available: https://keithito.com/LI-Speech-Dataset/. [Accessed: 20-May-2022].
[15] Nal Kalchbrenner, et al., “Efficient neural audio synthesis,” Proc. of PMLR, vol.80 , pp 2410-2419, 2018.
[16] G. E. Karniadakis, I. G. Kevrekidis, L. Lu, P. Perdikaris, S. Wang, and L. Yang, “Physics-informed machine learning,” Nat. Rev. Phys., vol. 3, no. 6, pp. 422-440, 2021.
[17] P. Singh, M. K. Mukul, and R. Prasad, “Bone conducted speech signal enhancement using LPC and MFCC,” Hum. Comput. Interact., pp. 148-158, 2018.
[18] Y. Luo and N. Mesgarani, “Conv-TasNet: Surpassing ideal time-frequency magnitude masking for speech separation,” IEEE/ACM Trans. Audio Speech Lang. Process., vol. 27, no. 8, pp. 1256-1266, 2019.
[19] J.-M. Valin and J. Skoglund, “LPCNET: Improving neural speech synthesis through linear prediction,” Proc. of IEEE ICASSP, pp. 5891–5895, 2019.
[20] R. de Frein, “Power-weighted LPC formant estimation,” IEEE Trans. Circuits Syst. II, vol. 68, no. 6, pp. 2207-2221, 2021.
[21] T. Mellalhi and R. Hamdi, “LPC-based formant enhancement method in Kalman filtering for speech enhancement,” Int. J. Electron. Commun., vol. 69, no. 2, pp. 545-554, 2015.
[22] ITU, “P.862.2:Wideband extension to Recommendation P.862 for the assessment of wideband telephone networks and speech codecs,” Itu.int. [Online]. Available: https://www.itu.int/rec/T-REC-P.862.2. [Accessed: 20-May-2022].
[23] J. Shen et al., “Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions,” Proc. of IEEE ICASSP, pp.4779-4783, 2018.
[24] X. Zheng, B. Aragam, P. Ravikumar, and E. P. Xing, “DAGs with NO TEARS: Continuous optimization for structure learning,” Proc. of NeurIPS, 2018.
[25] R. A. Chiea, M. H. Costa, and G. Barraut, “New insights on the optimality of parameterized Wiener filters for speech enhancement applications,” Speech Commun., vol. 109, pp. 46-54, 2019.
[26] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, “An algorithm for intelligibility prediction of time-frequency weighted noisy speech,” IEEE Trans. Audio Speech Lang. Processing, vol. 19, no. 7, pp. 2125-2136, 2011.
[27] Y.-Y. Yang et al., “Torchaudio: Building Blocks for Audio and Speech Processing,” Proc. of IEEE ICASSP, 2022.
[28] D. Yin, C. Luo, Z. Xiong, and W. Zeng, “PHASEN: A phase-and-harmonics-aware speech enhancement network,” Proc. of Conf. AAAI Artif. Intell., vol. 34, no. 05, pp. 9438-9465, 2020.
[29] H. Zhao, S. Zarar, I. Tashev, and C.-H. Lee, “Convolutional-recurrent neural networks for speech enhancement,” Proc. of IEEE ICASSP, pp. 2401-2405, 2018.
[30] Andersen, N., “On the calculation of filter coefficients for maximum entropy spectral analysis,” Geophysics vol. 39, no. 1, pp. 69-72, 1974.
[31] Boersma, Paul, “Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound,” Proc. of the Institute of Phonetic Sciences, vol. 17, no. 1193, pp. 97-110, 1993.