Speech Super-Resolution Using Parallel WaveNet

Mu Wang[1]*, Zhiyong Wu*[12]*, Shiyin Kang*[3]*, Xixin Wu*[2]*, Jia Jia*[1]*, Dan Su*[3]*, Dong Yu*[3]*, Helen Meng*[12]

[1]Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems, Graduate School at Shenzhen, Tsinghua University, Shenzhen
[2]Department of Systems Engineering and Engineering Management, The Chiense University of Hong Kong, Hong Kong
[3]Tencent AI Lab, Shenzhen

wangmu16@mails.tsinghua.edu.edu, {zywu, wuxx, hmmeng}@se.cuhk.edu.hk, {shiyinkang, dansu, dyu}@tencent.com, jjia@tsinghua.edu.cn

Abstract

Audio super-resolution is the task to increase the sampling rate of a given low-resolution (i.e. low sampling rate) audio. One of the most popular approaches for audio super-resolution is to minimize the squared Euclidean distance between the reconstructed signal and the high sampling rate signal in a point-wise manner. However, such approach has intrinsic limitations, such as the regression to mean problem. In this work, we introduce a novel auto-regressive method for the speech super-resolution task, which utilizes WaveNet to model the distribution of the target high-resolution signal conditioned on the log-scale mel-spectrogram of the low-resolution signal. As an auto-regressive neural network, WaveNet uses the negative log-likelihood as the objective function, which is much more suitable for highly stochastic process such as speech waveform, instead of the Euclidean distance. We also train a parallel WaveNet to speed up the generating process to real-time. In the experiments, we perform speech super-resolution by increasing the sampling rate from 4kHz to 16kHz on the VCTK corpus. The proposed method can achieve the improvement of $\sim 2dB$ over the baseline deep residual convolutional neural network (CNN) under the Log-Spectral Distance (LSD) metric.

Index Terms: speech super resolution, bandwidth extension, Wavenet, parallel Wavenet, auto-regressive model

1. Introduction

Speech super-resolution is an inverse problem trying to increase the temporal resolution of speech signals. In frequency domain, speech super-resolution is also known as bandwidth extension with the purpose to restore the high-frequency part of the speech signal from the distorted observation with low sampling rate. Speech super-resolution has applications in many fields, such as telephony communication, speech audio compression and text-to-speech synthesis [1].

Recently, data-driven approaches, especially deep neural networks (DNNs), are very popular in lots of fields including inverse problems. Because super-resolution has little dependence on very far contexts, convolutional neural networks (CNNs) are frequently used in both image and audio super-resolution task [1, 2, 3]. Besides fast training, CNNs also provide stability to small geometric deformations and provide features with smaller variance [2]. The most popular objective used in these models is the (squared) Euclidean distance [1, 3, 4]. However, for super-resolution task, the mapping from the distorted observation to the ground truth is highly unstable and even multi-valued. The Euclidean distance objective often leads to low perceptual quality, because it allways tries to resgress to mean. Auto-regressive models can relieve this problem by condtioning the output on the past outputs. Moreover, using a distribution to model the target data gives us a variable variance over time instead of a constant one [5], which is important to model raw time-domain signals such as speech data.

Due to the above reason, in this paper, we propose the use of WaveNet model [6] as the auto-regressive model for the speech super-resolution task, where WaveNet is used to predict the high-resolution signal while being conditioned on the distorted audio signal. For fast sampling, we use the parallel version of WaveNet [7]. We also evaluate two kinds of conditioning approaches for the WaveNet, conditioning the sampling process on the distorted audio signal directly, or on the log-scale mel-spectrogram of the distorted observation. Experiments indicate that the former approach would generate lots of noise, while the latter one can generate much better result. The results show the feasibility and effectiveness of our approach.

2. Related Work

For a speech signal, its high-frequency part is considered highly dependent on its corresponding low-frequency counterpart. Speech super-resolution tries to recover such high-frequency part from the low-frequency observations of the low resolution audio samples. To construct the zoomed-up samples, deep CNNs are used by directly conditioning on the distorted version in time domain [1]. A general mean squared error (MSE) loss is used to perform the point-wise estimation [1]. Frequency domain modeling is also adopted to get better results [4], where spectral fusion is used to reconstruct signals by retaining magnitude from frequency branch and using phase from time branch. Some researchers use a custom SampleRNN-like model to do this task [8]. However, to make sure real-time synthesis, their model is not auto-regressive. For image super-resolution task, instead of using a naive point-wise MSE loss, researchers also propose to use as conditional model a Gibbs distribution, where its sufficient statistics are given by deep CNNs [2].

To avoid the intrinsic limitations of the Euclidean distance objective, auto-regressive models can be used. WaveNet is a high-quality neural vocoder, which is a typical auto-regressive model directly modeling raw signals in the time domain [6]. In the original paper, they use linguistic feature and fundamental frequency as local condition to synthesize meaningful audios. Recently, log-scale mel-spectrogram is found a good acoustic feature to be the local condition of WaveNet [9]. In spite of the high-quality of synthesized audio, the sampling process is computationally expensive, resulting in a very slow generating speed. Excitingly, a parallel version of WaveNet is proposed to speed up the generating process 2000x faster in a single GPU [7]. In this work, we use parallel WaveNet to perform speech super-resolution.

3. Approach

The general super-resolution task aims to estimate a high-dimensional vector $\overrightarrow{y} \in \mathbb{R}^H$ given a distorted low-resolution observation $\overrightarrow{x} \in \mathbb{R}^L$. Common mapping functions have the following form:

$$U\colon \overrightarrow{x} \mapsto \overrightarrow{y} \tag{1}$$

No matter what model chosed as the mapping function, that mapping form often leads to a point-wise Euclidean distantance objective [1, 3, 4]:

$$\mathcal{L} = \frac{1}{NH}\sum_{i=1}^{N}\sum_{j=1}^{H}(y_{i,j} - U(\overrightarrow{x_i})_j)^2 \tag{2}$$

where N is the number of samples in the training set.

Moreover, since the high-dimensional vector $\overrightarrow{y}$ only conditions on the low-dimensional observation $\overrightarrow{x}$, that mapping is highly unstable and even multi-valued [2].

In this work, we consider using auto-regressive models to do speech super-resolution.

3.1. Speech Super-Resolution with WaveNet

The auto-regressive mapping functions have a totally different form. For simplicity, we represent the mapping as a probability density function:

$$p(\overrightarrow{x}) = \prod_{t=1}^{T} p(x_t|\overrightarrow{x}_{<t}) \tag{3}$$

For speech super-resolution task, this equation can be represented more specifically:

$$p(\overrightarrow{y}|\overrightarrow{x}) = \prod_{t=1}^{T} p(y_t|\overrightarrow{y}_{<t}, \overrightarrow{x}) \tag{4}$$

where $\overrightarrow{x}$ stands for the low-resolution audio, $\overrightarrow{y}$ the high-resolution counterpart.

A common practice for this task is to predict the high-frequency residuals ($\overrightarrow{y}_{res}$) using the raw low-frequency signals ($\overrightarrow{x}$) [1, 2]:

$$p(\overrightarrow{y}_{res}|\overrightarrow{x}) = \prod_{t=1}^{T} p(y_{res,t}|\overrightarrow{y}_{res,<t}, \overrightarrow{x}) \tag{5}$$

We also try to directly predict the high-resolution audio by conditioning WaveNet on the log-scale mel-spectrogram of the low-resolution version, since mel-spectrogram is proved to be a good intermediate feature representation for WaveNet to generate high-fidelity audios [5]. The approach can be formulated as follows, where $\phi(\overrightarrow{x})$ is the log-scale mel-spectrogram of low-resolution signal $\overrightarrow{x}$:

$$p(\overrightarrow{y}|\phi(\overrightarrow{x})) = \prod_{t=1}^{T} p(y_t|\overrightarrow{y}_{<t}, \phi(\overrightarrow{x})) \tag{6}$$

$$\phi(\overrightarrow{x}) = \ln Melfilter(|STFT(\overrightarrow{x})|) \tag{7}$$

Naively sampling from the auto-regressive model is unacceptably time-consuming, since only one sample can be generated at each sampling step. Although there are many techniques to speed up this process, almost all of them can only *linear* speed up it. In this work, we use parallel WaveNet to speed up the sampling process:

$$p(y_t|\overrightarrow{z}_{<t}) = \mathbb{L}(y_t|\mu(\overrightarrow{z}_{<t}), s(\overrightarrow{z}_{<t})) \tag{8}$$

Equation 8 shows the logistic distribution outputted by the parallel WaveNet. $\mu(\overrightarrow{z}_{<t})$ and $s(\overrightarrow{z}_{<t})$ are Inverse Auto-Regressive Flows (IAFs). For simplicity, we omit the condition symbols and weights. $\overrightarrow{z}$ is drawn from $\mathbb{L}(0, I)$. Since y_t now is not conditioned on $\overrightarrow{y}_{<t}$, the sampling process can be parallelized.

The loss functions used in the original parallel WaveNet paper are the KullbackLeibler divergence (D_{KL}) between the student (P_S) and teacher (P_T) distribution, the power loss ($\mathcal{L}_{power}$) and two others. We found that using D_{KL} and $\mathcal{L}_{power}$ is sufficient to train a good parallel WaveNet model conditioned on log-scale mel-spectrogram.

$$D_{KL}(P_S||P_T) = H(P_S, P_T) - H(P_S) \tag{9}$$

To calculate the $H(P_S, P_T)$ loss term, in our implementation, we use the reparametrize trick, commonly used in Variational Auto-Encoders (VAEs), to generate lots of samples at each time step, and then evaluate them under the distribution outputted by the teacher WaveNet.

$$H(P_S, P_T) = \sum_{t=1}^{T} \mathbb{E}_{p_S(\overrightarrow{y}_{<t})} H(p_S(y_t|\overrightarrow{y}_{<t}), p_T(y_t|\overrightarrow{y}_{<t})) \tag{10}$$

$$= \sum_{t=1}^{T} \mathbb{E}_{p_S(\overrightarrow{y}_{<t})} \mathbb{E}_{p_S(y_t|\overrightarrow{y}_{<t})} - \ln p_T(y_t|\overrightarrow{y}_{<t}) \tag{11}$$

$$H(P_S) = \mathbb{E}_{z\sim\mathbb{L}(0,1)}[\sum_{t=1}^{T} \ln s(\overrightarrow{z}_{<t})] + 2T \tag{12}$$

$\mathcal{L}_{power}$ is very crucial to train a good student network. However, it is not an easy work to make good use of it to train the model.

Several forms of loss functions can be adopted to avoid the student from collapsing to a WaveNet mode [7], such as the (squared) Euclidean distance between $|STFT(\overrightarrow{y}_{gen})|$ and $|STFT(\overrightarrow{y}_{gt})|$, and the distance between the log-scale spectrogram. Whereas, although these loss functions are much more stable, they would decrease the quality of the synthesized audio. We also tried to preprocess the raw signal by μ-law and pre-emphasis, however, both of them would also decrease the quality of the generated audio. Hence, in this work, we used the power loss proposed in the parallel WaveNet paper [7], and didnt average $\psi(\overrightarrow{y})$ over time before taking the Euclidean distance:

$$\mathcal{L}_{power} = ||\psi(\overrightarrow{y}_{gen}) - \psi(\overrightarrow{y}_{gt})||_2^2 \tag{13}$$

$$\psi(\overrightarrow{y}) = |STFT(\overrightarrow{y})|^2 \tag{14}$$

where $\overrightarrow{y}_{gen}$ stands for the generated high-resolution audio, $\overrightarrow{y}_{gt}$ the ground-truth.

3.2. Model Specifications

The standard setup of the WaveNet teacher consists of 30 layers, grouped into 3 dilated residual blocks of 10 layers. To speed up convergence, we scale the waveform targets by a factor of 127.5 as suggested in [5]. The learning rate was set as 2×10^{-4} and never changed during training. The teacher network was trained for 140,000 steps with Adam optimizer [10], with a minibatch size of 4 audio clips, each containing 10,000 timesteps. When training the teacher, we added uniform noise to dequantize the audio signals as suggested in [11].

The WaveNet student consists of 4 flows with 10, 10, 10, 30 layers respectively. We generated 100 samples per timestep to calculate the $H(P_S, P_T)$ loss term. To calculate $\mathcal{L}_{power}$, the number of FFT dots and frame shift length we used are 1024 and 272 respectively. We also scale the target of the student by a factor of 127.5, which we found much helpful to speed up convergence. The student network was trained for 200,000 steps.

The spectrogram is calculated with Hann windowing, with the window length of 800, shift length of 200, and FFT dots of 1024. A two-layer, with strides of 10 and 20 respectively, transposed convolutional neural network is used to map the log-scale mel-spectrogram to a new time series with the same resolution as the audio signal [5].

4. Experiments

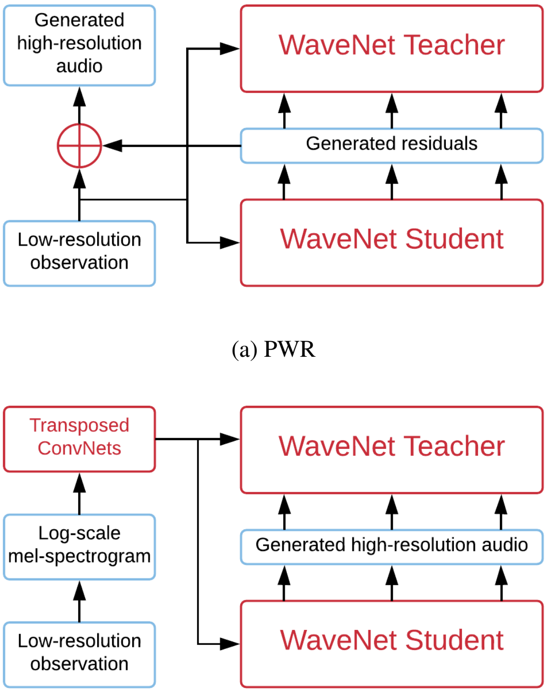
4.1. Datasets

We evaluate our approach with single-speaker speech super-resolution task. We chose 2 speakers, a female and a male, from the VCTK dataset [12] for experiments. The speech data for each speaker are about 30 minutes. The sampling rate of the original speech signal is 48 kHz. We down-sampled the signal to 16kHz as the high-resolution ground-truth, and then further down-sampled the signal to 4kHz as the low-resolution observation. We tested our model on 8 held-out records for each speaker.

4.2. Comparing Methods

In the conducted experiments, four methods for speech super-resolution are compared, including two baselines and two proposed WaveNet based approaches with different setups.

- Spline: The method increases the temporal resolution by cubic B-spline interpolation of waveform samples in the time domain.
- DRCNN: A deep residual CNNs used for audio super-resolution proposed in [1]. We instantiated the model with 4 down-sampling blocks and 4 up-sampling blocks. The model was trained for 120 epochs with Adam optimizer on audio clips, each containing 6,000 timesteps. The learning rate was 1×10^{-4}. We used the codes provided by the authors[1].
- PWR: Parallel WaveNet for speech super-resolution. It's conditioned on raw low-resolution audios and predicts residuals.
- PWM: Parallel WaveNet for speech super-resolution. It's conditioned on mel-spectrograms of low-resolution signal and directly predicts high-resolution version.


Generated high-resolution audio
WaveNet Teacher
Generated residuals
Low-resolution observation
WaveNet Student
Transposed ConvNets
WaveNet Teacher
Log-scale mel-spectrogram
Generated high-resolution audio
Low-resolution observation
WaveNet Student

(a) PWR

(b) PWM

Figure 1: *Parallel WaveNet for speech super-resolution.*

4.3. Metrics

We use the Log-Spectral Distance (LSD) [13] to measure the reconstruction quality.

$$LSD(x, y) = \frac{1}{L} \sum_{l=1}^{L} \sqrt{\frac{1}{K} \sum_{k=1}^{K} (X_{gen}(l,k) - X_{gt}(l,k))^2} \quad (15)$$

$$X = \ln |STFT|^2 \quad (16)$$

l is used to index frames, and k is to index frequencies. We used a frame length of 2048 to calculate the LSD.

Table 1: *LSD evaluation of audio-super resolution methods (in dB) at upscaling ratios r = 4 from sampling rate 4kHz to 16kHz.*

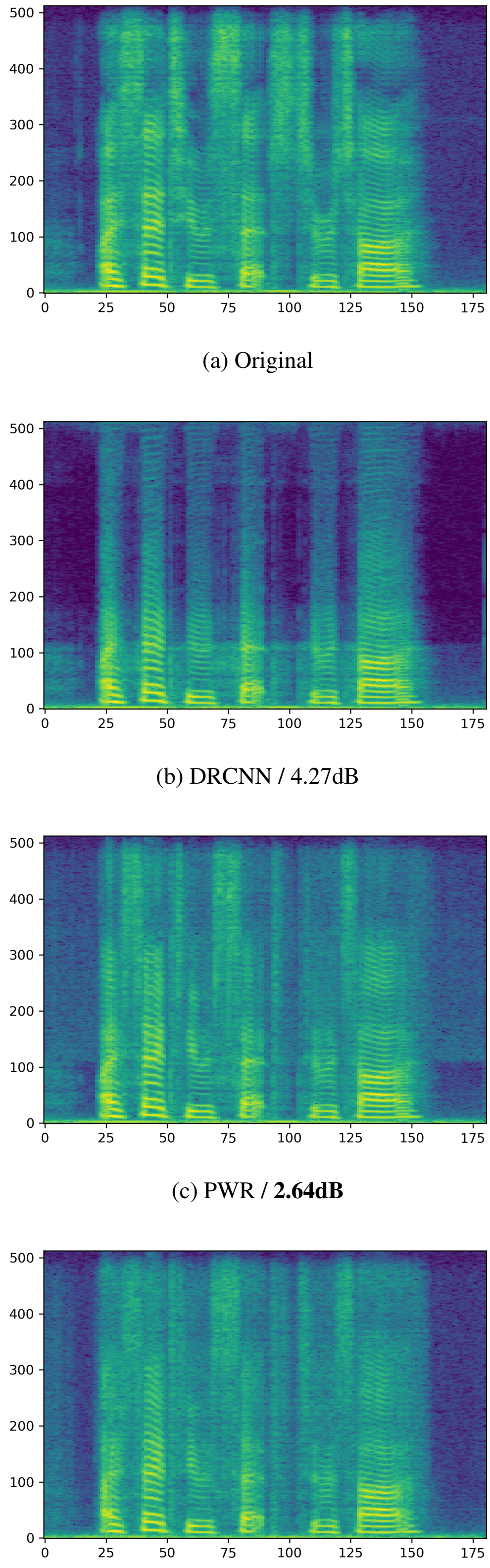
	Female (p225)	**Male** (p226)
Spline	8.41	8.46
DRCNN	4.03	4.52
PWR	**2.51**	**2.38**
PWM	**2.14**	**2.10**

4.4. Results and Analysis

The LSD evaluation results of the four comparing methods are illustrated in Table. 1. As can be seen, our proposed PWM method shows an improvement of $\sim 2dB$ over the DRCNN model proposed in [1]. More audio samples are available [2] for

[1] https://github.com/kuleshov/audio-super-res

[2] Audio samples are available at http://t.cn/RB539Wp



(a) Original

(b) DRCNN / 4.27dB

(c) PWR / **2.64dB**

(d) PWM / **2.17dB**

Figure 2: *Log-scale spectrograms of super-resolution examples generated with different methods (b, c, d) and the original high-resolution signal (a).*

listening. The log-scale spectrograms of examples generated by different methods are shown in Fig. 2.

When listening to these generated audios, we found that the audios generated by the baseline neural network lacked the high-frequency part of the original high-resolution version, although the high-frequency part of their spectrogram seems normal. Besides that problem, the generated audios had artificial noises. The possible explanation is that the very high-frequency part of human speech is highly stochastic, although some missing high-frequency bands have a much more deterministic correlation to the low-frequency bands. Since the standard Euclidean distance is suffered from the *regression to mean* problem, such highly stochastic signals are very hard to predict using that objective function.

Although the audios generated by parallel WaveNet are better than the baselines, there are still background noises. In our experiments, we found that directly conditioning the generating process on the (bicubic-upsampled) raw distorted audios (i.e. the PWR method) would generate lots of noises, while using mel-spectrogram as local condition (i.e. the PWM method) gave us much cleaner audios. A possible explanation is that the STFT function is just like a non-trainable convolutional network with a big reception field. By using spectrogram as the local condition, the generating process could even use the future information of the low-resolution signals. We suspect that a trainable CNN module with a big reception field may perform equally well and even better.

5. Conclusions

In this paper, we demonstrate the feasibility and effectiveness of using parallel WaveNet to do speech super-resolution task. Our method greatly outperforms the baselines under the LSD metric. We found that using spectrogram as the local condition of the WaveNet model is better than raw signal. In the future work, we plan to find a better local condition, for example, using a trainable CNN module to replace the STFT calculation, because spectrogram only has magnitude information of the distorted signal. The noise in the generated audio is a problem, we will try to tackle it in our future work.

6. Acknowledgement

This work is supported by joint research fund of National Natural Science Foundation of China - Research Grant Council of Hong Kong (NSFC-RGC) (61531166002, N_CUHK404/15), National Natural Science Foundation of China (61433018, 61375027) and National Social Science Foundation of China (13&ZD189). We would also like to thank Tencent AI Lab Rhino-Bird Joint Research Program (No.JR201803) and Tsinghua University - Tencent Joint Laboratory for the support.

7. References

[1] V. Kuleshov, S. Z. Enam, and S. Ermon, “Audio super resolution using neural networks,” *The workshop track of the International Conference on Learning Representations (ICLR)*, 2017.

[2] J. Bruna, P. Sprechmann, and Y. LeCun, “Super-resolution with deep convolutional sufficient statistics,” *arXiv preprint arXiv:1511.05666*, 2015.

[3] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, “Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1874–1883.

[4] T. Y. Lim, R. Yeh, Y. Xu, M. N. Do, and M. Hasegawa-Johnson, “Time-frequency networks for audio super-resolution,” *Proceedings of the 43nd International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[5] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerry-Ryan *et al.*, “Natural tts synthesis by conditioning wavenet on mel spectrogram predictions,” *Proceedings of the 43nd International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[6] A. Van Den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, “Wavenet: A generative model for raw audio,” *arXiv preprint arXiv:1609.03499*, 2016.

[7] A. v. d. Oord, Y. Li, I. Babuschkin, K. Simonyan, O. Vinyals, K. Kavukcuoglu, G. v. d. Driessche, E. Lockhart, L. C. Cobo, F. Stimberg *et al.*, “Parallel wavenet: Fast high-fidelity speech synthesis,” *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2018.

[8] Z.-H. Ling, Y. Ai, Y. Gu, and L.-R. Dai, “Waveform modeling and generation using hierarchical recurrent neural networks for speech bandwidth extension,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 5, pp. 883–894, 2018.

[9] A. Gibiansky, S. Arik, G. Diamos, J. Miller, K. Peng, W. Ping, J. Raiman, and Y. Zhou, “Deep voice 2: Multi-speaker neural text-to-speech,” in *Advances in Neural Information Processing Systems*, 2017, pp. 2966–2974.

[10] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.

[11] L. Theis, A. v. d. Oord, and M. Bethge, “A note on the evaluation of generative models,” *arXiv preprint arXiv:1511.01844*, 2015.

[12] J. Yamagishi, “English multi-speaker corpus for cstr voice cloning toolkit,” *URL http://homepages. inf. ed. ac. uk/jyamagis/page3/page58/page58. html*, 2012.

[13] A. Gray and J. Markel, “Distance measures for speech processing,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 5, pp. 380–391, 1976.