EURASIP Journal on Advances in Signal Processing

RESEARCH Open Access

A multi-task learning speech synthesis optimization method based on CWT: a case study of Tacotron2

Guoqiang Hu[1], Zhuofan Ruan[2], Wenqiu Guo[3] and Yujuan Quan[4*]

*Correspondence:
tquanyj@jnu.edu.cn

[1] International School, Jinan University, No. 855 Xingye Avenue East, Guangzhou 511486, Guangdong, China
[2] Information Hub, The HONG KONG University of Science and Technology(Guangzhou), No.1 Duxue Road, Guangzhou 511453, Guangdong, China
[3] School of Business, Macau University of Science and Technology, Avenida Wai Long, Macao 999078, China
[4] College of Information Science and Technology, Jinan University, No. 855 Xingye Avenue East, Guangzhou 511486, Guangdong, China

Abstract

Text-to-speech synthesis plays an essential role in facilitating human-computer interaction. Currently, the predominant approach in Text-to-speech acoustic models selects only the Mel spectrum as an intermediate feature for converting text to speech. However, the Mel spectrograms obtained may exhibit ambiguity in some aspects owing to the limited capability of the Fourier transform to capture mutation signals during the acquisition of the Mel spectrograms. With the aim of improving the clarity of synthesized speech, this study proposes a multi-task learning optimization method and conducts experiments on the Tacotron2 speech synthesis system to demonstrate the effectiveness of the proposed method. The method in the study introduces an additional task: wavelet spectrograms. The continuous wavelet transform has gained significant popularity in various applications, including speech enhancement and speech recognition, which is primarily attributed to its capability to adaptively vary the time-frequency resolution and its excellent performance in capturing non-stationary signals. This study highlights that the clarity of Tacotron2 synthesized speech can be improved by introducing Wavelet-spectrogram as an auxiliary task through theoretical and experimental analysis: a feature extraction network is added, and Wavelet-spectrogram features are extracted from the Mel spectrum output generated by the decoder. Experimental findings indicate that the Mean Opinion Score achieved for the speech synthesized by the model using multi-task learning is 0.17 higher compared to the baseline model. Furthermore, by analyzing the factors contributing to the success of the continuous wavelet transform-based multi-task learning method in the Tacotron2 model, as well as the effectiveness of multi-task learning, the study conjectures that the proposed method has the potential to enhance the performance of other acoustic models.

Keywords: Speech synthesis, Continuous wavelet transform, Multi-task learning, Clarity, Tacotron2

1 Introduction

Research in the field of speech synthesis [1, 2] has its origins in the 17th and 18th centuries, during which machine devices were employed to generate sounds that imitated the human vocal organs. Over an extended period of time, several approaches have emerged in the field of speech synthesis. These include resonance peak parameter-based speech synthesis, waveform concatenation-based speech synthesis, and statistical parameter-based speech synthesis. However, parameter-based speech synthesis with resonant spikes necessitates a substantial volume of training data; otherwise, the effectiveness of synthesis is diminished. In addition, it is imperative to train separate models for different speakers in order to maintain the quality of the synthesized speech. Simultaneously, it also exhibits subpar performance in certain facets of speech intricacies. Waveform concatenation-based speech synthesis and statistical parameter-based speech synthesis are both associated with certain challenges. The former is known for difficulties in achieving smooth speech splicing and is prone to producing discontinuity or noise. On the other hand, the latter requires significant computational resources, exhibits poor generalization ability, and has a low synthesis speed.

With the popularity of deep learning, text-to-speech synthesis models [3] based on deep neural networks have gradually become mainstream due to their superior speech synthesis performance and strong generalization ability. These models mainly consist of three parts: a text frontend, an acoustic model, and a vocoder. Acoustic models play a crucial role in speech synthesis by converting text into the corresponding acoustic features. In recent years, there has been a surge in the development of neural network-based acoustic models that have demonstrated superior performance and distinct characteristics. Examples of such models include Tacotron2 [4], Transformer TTS [5] and Fastspeech [6]. These models have significantly contributed to the advancement of speech synthesis.

Acoustic features can be obtained as intermediate features through various signal processing methods, which play a crucial role in Text to speech. The most dominant signal processing methods are Fourier transform and wavelet transform. The wavelet transform [7] is extensively utilized in signal processing [8] and image analysis [9] due to its several variations, including the continuous wavelet transform(CWT), discrete wavelet transform [10, 11], fractal-wavelet analysis [12, 13], Chebyshev wavelet analysis [14], among others, as per the algorithmic approach.

The majority of acoustic models choose Mel only as an intermediate text-to-speech feature because Mel better mimics the perceptual characteristics of the human ear. However, Mel is limited by Short-time Fourier transform [15], which employs a fixed-width window, so the resolution at various frequencies remains constant, resulting in capturing the time-frequency characteristics of speech signals. In addition to this, the Short-time Fourier transforms exhibit a relatively moderate response to variations in the speech signal, necessitating the use of longer time windows to effectively capture these variations. Based on the aforementioned limitations of Short-time Fourier transform, the dynamic window size of CWT [16] offers a more effective approach for analyzing instantaneous variations of speech signals. By employing a dynamic window size, the continuous wavelet transform is able to provide varying resolutions at different frequencies, thereby capturing the time-frequency characteristics of speech signals more

accurately. In this study, we propose a novel approach that leverages multi-task learning [17] to integrate the spectrogram obtained by continuous wavelet transform and the Mel spectrogram obtained through Short-time Fourier transform as acoustic features [18]. Our objective is to establish the connection between text, Mel spectrogram, and wavelet spectrogram. The Mel is employed as the main training objective, while the wavelet spectrogram serves as an auxiliary objective. The objective of improving the performance of the obtained Mel spectra is thus accomplished. Moreover, there have been subsequent improvements in the clarity of speech generated by Text-to-Speech systems and in their ability to capture mutation signals.

The paper includes the following sections. The first part introduces the techniques used to acquire Mel from different mainstream acoustic models, as well as the CWT introduced in Fastspeech2 [19] for extracting pitch spectra [20, 21]. Section II presents the definition of multi-task learning along with its core and auxiliary tasks. Section III provides a comprehensive explanation of the CWT and its application in processing the wav file to acquire the wavelet spectrum. Section IV introduces the model architecture of Tacotron2, incorporating CwtNet, along with the data processing techniques employed and the experimental results obtained during the conducted experiments. Section V provides an analysis of the strengths and weaknesses of the research methodology employed in this paper. Additionally, it explores future directions for enhancing the effectiveness of alternative acoustic models.

2 Related acoustic model

Since the hot research on deep learning-based speech synthesis [22], most mainstream acoustic models have completed text-to-Mel translation directly through neural networks. Work on per-acoustic models has been devoted to obtaining more accurate Mel correspondences with the text through different model architectures, which can be categorized into basic models such as CNN [23], RNN [24], and Transformer [25]. For example, Baidu's Deepvoice [26–28] series uses a fully convolutional encoder–decoder architecture to transform text to Mel through causal convolutional blocks and attention blocks. The optimization objective of this paper, Tacotron2, mainly uses LSTM to generate Mel for each frame in a sequential manner. Transformer TTS uses a Transformer architecture whose output is processed by Post-net to obtain the Mel corresponding to the text.

Mel has been consistently used as the traditional training objective during the training of the aforementioned acoustic models [29], and CWT has not played a significant role in speech synthesis at this stage. However, in 2020, Fastspeech2, jointly published by Zhejiang University and Microsoft, considerably improved the performance of synthesized speech by modeling duration, pitch, and energy, and its fundamental frequency predictor applied CWT, which obtained the fundamental frequency spectrogram by CWT on continuous fundamental frequency sequences and better predicted the contour changes of the fundamental frequency. In inference, the fundamental frequency predictor predicts the fundamental frequency spectrogram and further converts it back to the fundamental frequency contour by using continuous wavelet inverse transform (iCWT).

Fastspeech2 uses the characteristics of CWT to improve the performance of synthesized speech, which also proves that the advantage of wavelet transforms over short-time Fourier transform (which will be discussed in detail in Part 3) can be used as a new direction to improve speech synthesis performance [30]. Inspired by this approach, the main work of this paper is to exploit the advantages of CWT from a different perspective.

3 Multi-task learning

Definition: A highly useful machine learning strategy that aims to improve model generalization and performance by jointly learning several different but related tasks [31]. The main task generally shares a part of the representation with additional auxiliary tasks, and auxiliary tasks can contribute to the model training of the corresponding main task by supplementing information, transferring knowledge, and increasing the amount of training data [31]. The model framework based on multi-task learning is shown in Fig. 1 [32], and its Composite loss function is given below in Eq. (1).

$$\epsilon_{\mathrm{MTL}} = \sum_{k=1}^{K} w_k \epsilon_k \tag{1}$$

MTL methods can be easily deployed on stochastic neural networks by sharing certain hidden layers between different tasks. MTL has made great achievements in several speech signal processing fields such as speech synthesis [31, 33, 34] and automatic speech recognition [35]. Whereas speech-related tasks typically involve various complex metrics, multi-task learning can consider multiple objective functions simultaneously and solve the problem by finding a set of optimal solutions that balance these multi-tasks. In the field of speech recognition and speech enhancement, numerous studies use multi-objective optimization to improve the performance of their models. Peng et al. [36] proposed a multi-objective speech enhancement model based on perceptual features. By modeling the speaker, pitch, and energy separately, it extends the original MSE loss for individual LPS features to an objective loss for pitch and speaker identity and obtains a composite loss function. After the experimental validation, the evaluation metrics of the models using the composite loss function are compared with the original ones. There is a certain amount of improvement.

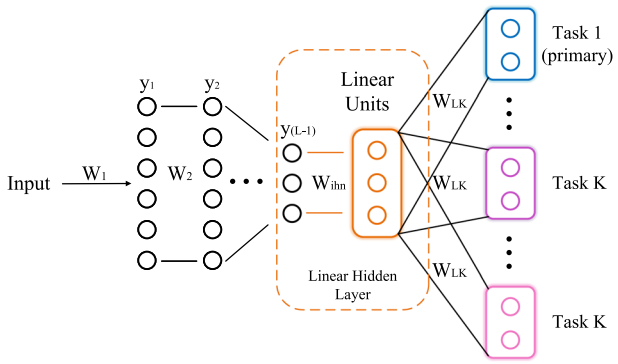

Fig. 1 Basic framework for multi-task learning

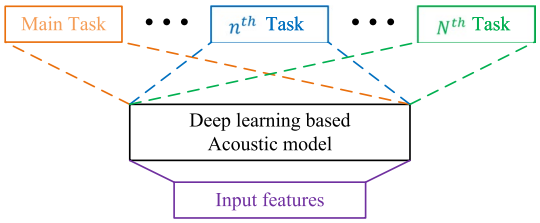

Fig. 2 Deep learning acoustic models with multi-task learning

In the field of speech synthesis, there is the famous one-to-many problem [37], that is, a text can correspond to multiple speech expressions. In Fastspeech2, a variance adaptor is used to predict pitch, energy, and duration separately, which partially solves this problem. It is fundamentally similar to multi-task learning, which relies on auxiliary tasks to complement information and transfer knowledge. The application framework of multi-task learning in the field of speech synthesis is shown in Fig. 2 [38] below. How to provide additional prior knowledge to our model so that it can generate speech that is more in line with our expectations in a myriad of one-to-many problems is a mainstream direction in the field of speech synthesis.

3.1 Core and auxiliary tasks: Mel spectrogram and wavelet spectrogram

Several studies have shown that humans are extra sensitive to nonlinear sounds. The Mel spectrogram represents the frequency distribution of the sound signal in the Mel scale, a nonlinear scale [39] that better models the perceptual properties of the human ear. As a result, the Mel spectrogram better reflects the perceptual properties of the human auditory system and is more consistent with human auditory characteristics. This is why most mainstream acoustic models choose Mel as the acoustic feature for synthesizing speech waveforms. The Mel spectrum is obtained by applying operations such as the short-time Fourier transform to the speech waveform [40], and both CWT and short-time Fourier transform are signal processing methods with their own advantages and disadvantages.

As mentioned earlier, the short-time Fourier transform is limited by its fixed window length and cannot balance time and frequency. In contrast, the CWT can provide time-frequency analysis [41] with arbitrary time and frequency resolution, thus better capturing local features such as non-stationary parts in the signal (a more detailed explanation of CWT will be given in Section IV). Therefore, we use the efficient capturing ability of the CWT for non-stationary signals [42] in speech to reduce the error in time-frequency analysis caused by obtaining the Mel spectrum from short-time Fourier transform(STFT) and to enhance the comprehensiveness of Tacotron2.

In summary, we believe that joint training [43] of Mel and continuous wavelet transform-based spectrograms can help neural network models to more comprehensively capture both global and local features of stationary signals, leading to a Mel that is more consistent with our expectations.

4 Continuous wavelet transform

CWT is a signal processing [44] method that provides a more interpretable visual representation of signals. It performs a time-frequency analysis on a signal by scaling and shifting a wavelet-generating function. For a signal $f(t)$, its CWT result is shown in Eq. (2):

$$\Psi f(a,b) = \int_{-\infty}^{+\infty} f(t)\overline{\psi}_{a,b}(t)\mathrm{d}t = \int_{-\infty}^{+\infty} f(t)\exp\left(-i\omega_0\frac{(t-b)}{a}\right)\exp\left(-\frac{(t-b)^2}{2a^2}\right)\mathrm{d}t \tag{2}$$

where a represents the scale factor, b represents the time shift factor, $f(t)$ is the original speech signal, and $\psi a, b(t)$ is the wavelet mother function.

Taking the Morlet wavelet as an example, the mathematical expression is as in Eq. (3) (note that CWT has multiple wavelet bases, and the appropriate wavelet basis should be selected based on the characteristics of the processed signal for better results).

$$\psi(t) = \exp(i\omega_0 t)\exp\left(-\frac{t^2}{2}\right) \tag{3}$$

The steps of the continuous wavelet transform are as follows.

1. Compare the starting part of the wavelet $w(t)$ and the original function $f(t)$ (essentially taking the inner product) and calculate the coefficient C. The coefficient C represents the similarity between the part of the function and the wavelet.
2. Shift the wavelet to the right by k units to obtain the wavelet $w(t-k)$ and repeat step 1. Repeat this step until the function $f(t)$ is complete.
3. Expand the wavelet $w(t)$ to obtain the wavelet $w(\frac{t}{2})$, and repeat steps one and 2.
4. Continuously expand the wavelet and repeat steps 1, 2 and 3.

As can be seen from Definition and Eq. (2), the essence of the CWT is to decompose the signal into a sequence of primitive signals with excellent time-frequency localization. The two ends of the basis function decay rapidly to zero and can move in the whole time domain, and the basis function can be changed by transformations of the expansion factors; therefore, the CWT is a time-frequency analysis method with variable resolution in both time and frequency domains. It can use long time intervals to obtain more accurate low-frequency information and short-time intervals to obtain high-frequency information. This is in line with the characteristics that the low-frequency part of the signal changes slowly and the high-frequency part changes rapidly, so it is known as the "microscope of signal analysis". At the same time, the wavelet transform also has the following differences from the short-time Fourier transform:

1. Wavelet transform does not perform Fourier transform on the windowed signal, and the transformed signal behaves differently.
2. In the process of wavelet transform, the width of the "window function" can be changed for each spectrum calculation, which is the most significant difference between the two.

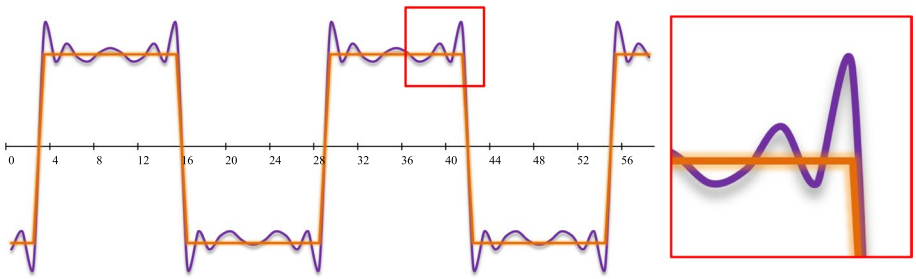

Fig. 3 Fourier transform fitting non-stationary signals

3. From the filter point of view, the bandwidth of the bandpass filter is independent of the center frequency for the short-time Fourier transform but proportional to the center frequency for the wavelet transform.

4.1 Gibbs effect [45]

Definition The Fourier transform has to be fitted with a large number of triangular wave functions for a signal with a sudden and violent transformation, even for a short period of time. As shown in Fig. 3, it illustrates the case of the Fourier transform when fitting a non-stationary signal.

Because its basis functions are trigonometric functions of infinite length, the Fourier transform has a poor ability to capture non-stationary signals. For a mutation signal, its spectrum usually contains more high-frequency components, so high-frequency resolution is needed to accurately represent its spectrum characteristics. Because of its fixed window length, Fourier transform cannot have good performance in both time resolution and frequency resolution, so it cannot capture the details of mutation in the spectrum.

The basis function of CWT is a finite and decaying wavelet function, which fits the signal by scaling and translating the wavelet function. According to its mathematical formula, the coefficient of the wavelet function is not equal to 0 only when it overlaps with the mutation signal, so that the mutation signal can be accurately captured. Through the CWT of a speech signal, the signal can be accurately divided into different timescales, so that the high-frequency components of the mutation signal can be better captured, so as to obtain a more accurate spectrum. Related papers show that CWT has been successfully applied to speech synthesis, voice conversion and other research [46].

4.2 Time-frequency resolution [47]

The short-time Fourier transform has a clear physical meaning and can give a time-frequency structure that matches our intuitive perception structure. However, due to the restriction of the uncertainty principle on the time-frequency resolution of the window function, that is, when the window length is overly long, its time resolution is poor, and when the window length is too short, its frequency domain resolution is poor. Therefore, short-time Fourier transform cannot obtain great resolution in both the time domain and

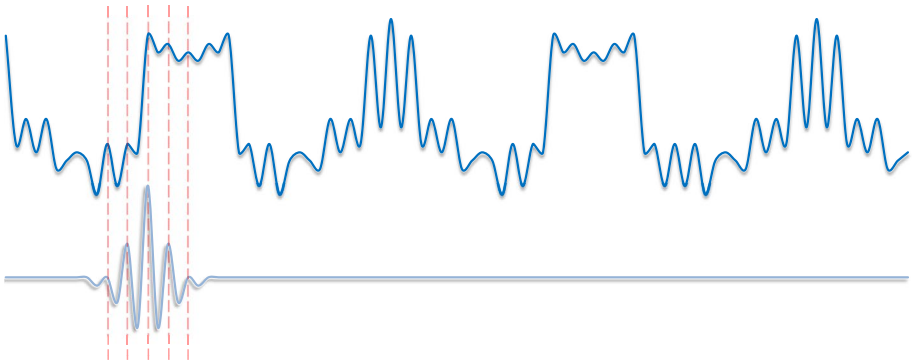

Fig. 4 Wavelet basis function fitting non-stationary signals through translation and stretching

frequency domain [48]. Compared with wavelet transform, wavelet transform is a more powerful signal analysis method, which is presented in Fig. 4, and its variable time-frequency resolution can meet the requirements of "seeing the forest for the trees" in signal processing. According to the characteristics of the signal, the time-frequency resolution can be adjusted adaptively to better capture the details in the signal spectrum, so as to obtain a more accurate spectrum. Compared with STFT, wavelet transform has multi-resolution characteristics [49].

5 Experiments and results

5.1 Dataset and preprocessing

The experiments in this paper use the English speech dataset LJspeech [50], which contains audio files of 13,100 sentences with a total of about 24 h of speech data with a sampling rate of 22,050 Hz.

In the data preprocessing phase, we perform signal processing methods such as normalization and short-time Fourier transform on each speech to obtain the true Mel spectrum corresponding to each speech as the core task of multi-task learning. At the same time, we also perform CWT on the speech and reduce the data dimension by Truncated SVD [51] and take the wavelet spectrum after dimension reduction as the auxiliary task of the multi-task learning.

Truncated Singular Value Decomposition (TSVD) is a dimension reduction method based on matrix decomposition. By performing Singular Value Decomposition (SVD) on the original data matrix, it obtains three essential components of the matrix: the left singular vector, the singular value, and the right singular vector. We can use these parts to reconstruct the original data matrix, but keep only the most salient parts, which are the first k singular values and the corresponding left and right singular vectors. This procedure is equivalent to the dimensionality reduction of the original data matrix, which reduces the dimensionality of the data while retaining the saliency information. In this paper, we use it to reduce the dimensionality of the sparse matrix after continuous wavelet transform to obtain a fixed dimension and non-sparse matrix, which will be beneficial for neural networks for feature learning.

5.2 Baseline Tacotron2

Tacotron2 consists of an encoder and a decoder with attention. The encoder transforms the character sequence into an implicit feature representation, which is used by the decoder to predict the spectrogram. Input characters are represented using learned 512-dimensional character em- beddings passed through a stack of three convolutional layers, each containing 512 filters of shape 5×1, that is, each filter spanning five characters, followed by batch normalization [52] and ReLU activation. These convolutional layers model long-term context (e.g., N-grams) in the input character sequence. The output of the final convolutional layer is passed into a single bidirectional [53] LSTM [54] layer with 512 units (256 in each direction) to generate the encoded features. Finally, the predicted Mel spectrogram is passed through a 5-layer post-convolutional network that adds a residual to the prediction to improve the overall reconstruction. Each postNetwork layer consists of 512 batch normalized filters of shape 5×1 with tanh activations on all but the last layer.

5.3 Proposed model architecture

Figure 5 shows the model architecture of our proposed multi-task learning tacotron2. Based on the original Tacotron2, the CwtNet feature extraction network is added to the model. For the frame-by-frame Mel spectra output by the decoder, CNN and fully connected neural networks are used to extract features to obtain wavelet spectra, that is, the frame-by-frame Mel spectra output by the decoder is required to perform multiple tasks. The core task is to use it as input to reconstruct and enhance the Mel spectrum through PostNet, and the auxiliary task is to use it as input to obtain the wavelet spectrum of the corresponding speech through CwtNet.

In the field of speech synthesis, numerous studies only minimize the mean square error between the predicted Mel and the true Mel. As shown in Eq. (4) below, in Tacotron2, its

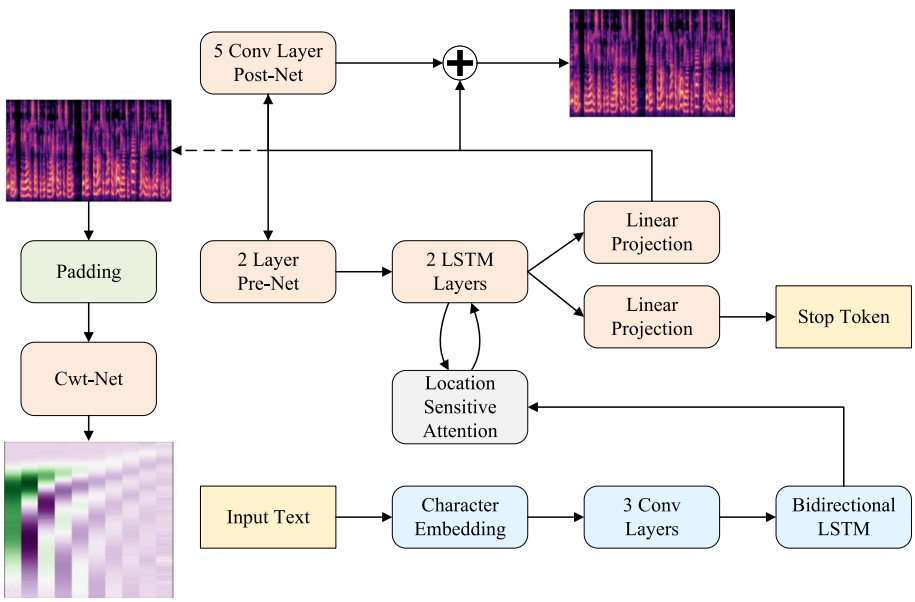

Fig. 5 The proposed multi-task learning Tacotron2 model

loss function consists of three parts, the decoder predicting the mean square error between

the merged Mel and the true Mel frame by frame, the mean square error between the PostNet reconstruction enhanced Mel and the true Mel, and the binary cross-entropy logic loss between the stopping probability and the true stopping probability of each frame.

$$\text{Loss}_{\text{Tacotron2}} = \frac{1}{n}\sum_{i=1}^{n}(y_1 - y)^2 + \frac{1}{n}\sum_{i=1}^{n}(y_2 - y)^2 + L_{\text{BCEwithlogitsloss}}(\hat{Z}, Z) \tag{4}$$

where y represents the true Mel spectrum, y_1 represents the Mel spectrum from the decoder output, y_2 represents the Mel spectrum from the PostNet output, and Z represents the stop token.

In our opinion, it is not sufficient to use only the acoustic feature Mel as the loss. Although the nonlinear features of Mel can be well adapted to the human cochlea, the Fourier transform in obtaining Mel has certain drawbacks in capturing non-stationary signals. A better approach would be to jointly train Mel and Wavelet spectrogram and apply appropriate weights to the Wavelet spectrogram loss so that the model can distinguish between core and auxiliary tasks in multi-task learning. Thus, we not only keep the original Mel loss of Tacotron2, but we also add a fourth loss function to predict the mean squared error between the predicted and true Wavelet spectrogram. Finally, we can formulate the composite loss function in multi-task learning for Tacotron2 as presented in Eqs. (5) and (6).

$$\text{Loss}_{\text{wavelet}} = \frac{1}{n}\sum_{i=1}^{n}(\hat{x} - x)^2 \tag{5}$$

$$\text{Loss}_{\text{MTL}-\text{Tacotron2}} = \text{Loss}_{\text{Tacotron2}} + \text{Loss}_{\text{wavelet}} \tag{6}$$

where x represents the true wavelet spectrogram after TSVD, and $\hat{x}$ represents the predicted wavelet spectrogram after TSVD.

5.4 Experimental setting

The official training configuration for the Tacotron2 is a V100 configuration with eight graphics cards. Due to the limited computational power, the experimental part of this paper will use the official public statedict of Baseline Tacotron2 after 500 epochs of training as the initial parameters. The training starts with default initial parameters. During training, we use CwtNet to fine-tune the model parameters in Tacotron2, while most of the training parameters are the same as in Tacotron2, except that we use 16 as the batch size. In addition, due to the large size of the wavelet spectrogram data, in order to prevent it from too much influence on the model's learning of Mel, away from its original purpose as an auxiliary task to help the model learn Mel, we apply 0.004 weight to its loss function (this parameter can be adjusted as needed), so as to ensure that it is numerically equivalent to initial Mel_{loss}. During training, models with partially shared parameters are optimized for training on Mel_{loss}.

5.5 Problems in Tacotron2

Tacotron2 is the first end-to-end speech synthesis model, but there are still some audio quality problems to be solved. We tested partial speech synthesis using Tacotron2 and found that it suffers from problems such as ambiguous articulation at the beginning and end, inaccurate intonation, ambiguous words near the sound, overly quick speech speed, and unnatural intonation. Moreover, our proposed Tacotron2 based on multi-task learning improves the model's learning of non-stationary signals in the speech by applying an auxiliary task wavelet spectrogram, which partially solves the aforementioned problems of Tacotron2 synthesized speech and improves the clarity of the synthesized speech. The corresponding test audio has been used for audio test experiments (please refer to experimental results for details).

5.6 Experimental result

We aim to evaluate the audio quality of the speech synthesized by each of the two models from the test set as well as from outside the dataset. We keep the content of the text input accepted by Tacotron2 and our proposed model consistent, thus excluding additional interference factors to focus on audio quality. For each audio, we invited at least 30 normal esthetic testers to listen to it, including foreigners, English major students, who have normal foreign listening ability in English. Experimental subjects are 20 English audio clips synthesized by Tacotron2 and our proposed Tacotron2 based on multi-task learning. We selected MOS(Mean Opinion Score), the Mainstream evaluation metric for speech, as the subjective evaluation metric, and respondents were asked to rate each audio segment on a scale of 0–5, with higher scores indicating better quality of the corresponding speech.

The experimental results are shown in Table 1, from which we can find that Tacotron2 based on multi-task learning evidently gains extra preferences from listeners. Compared to Tacotron2, its MOS is improved by 0.17, indicating that multi-task learning helps the speech synthesis model to improve the quality of its synthesized audio.

6 Discussion and future work

In this paper, we first propose a Tacotron2 speech synthesis model based on a multi-task learning optimization method. By adding an auxiliary task (model learning for wavelet spectrogram), we improve its learning effect compared with the baseline model on its core task (model learning for Mel spectrum). In the previous section, our experimental results have demonstrated the success of our proposed auxiliary task using CWT on Tacotron2. However, the key to the success of the proposed model is that the CWT compensates for the shortcomings of STFT's low sensitivity to abrupt signals and fixed time-frequency resolution, so for other mainstream acoustic models, such as Transformer TTS, which also has PostNet, we can also try to improve its effectiveness by adopting the

Table 1 Comparison of evaluated MOS for Tacotron2 versus MTL-Tacotron2

Model	MOS
Baseline	3.70
MTL-Tacotron2	3.87

idea of multi-task learning optimization. This is also the direction of our future research. However, it consumes a lot of computational resources due to the high computational complexity of the CWT. How to reduce the use of computational resources or use wavelet transform with lower time complexity is one of the optimization ideas for the proposed model.

Author contributions
Conceptualization, project administration, investigation, writing, writing original draft were contributed by G.H.; funding acquisition, writing—review and editing, were contributed by Y.Q. and Z.F.; methodology was contributed by Z.F. and W.G.; supervision, resources were contributed by W.G. and Y.Q. All authors have read and agreed to the published version of the manuscript.

Funding
This research was funded by the Guangdong Basic and Applied Basic Research Foundation (2021A1515011999), Guangdong Key Laboratory of Traditional Chinese Medicine Informatization, Guangdong Province Big Data Innovation Engineering Technology Research Center, "Outstanding Future" Data Scientist Incubation Project of Jinan University, and the Fundamental Research Funds for the Central Universities, Jinan University (21619412).

Availability of data and materials
The datasets generated and/or analyzed during the current study are available in the LJ Speech Dataset repository, https://keithito.com/LJ-Speech-Dataset/.

Declarations

Competing interests
The authors declare no competing interets.

Received: 19 September 2023 Accepted: 6 December 2023
Published online: 02 January 2024

References

1. H. Zen, T. Toda, An Overview of Nitech HMM-based Speech Synthesis System for Blizzard Challenge 2005, in *Proceeding of the Interspeech 2005*. ISCA, Lisbon, Portugal, pp. 93–96 (2005). https://doi.org/10.21437/interspeech.2005-76
2. N. Kaur, P. Singh, Conventional and contemporary approaches used in text to speech synthesis: a review. Artif. Intell. Rev. **2022**, 1–44 (2022). https://doi.org/10.1007/s10462-022-10315-0
3. Y. Ning, S. He, Z. Wu, C. Xing, L.-J. Zhang, A review of deep learning based speech synthesis. Appl. Sci. **9**, 4050 (2019). https://doi.org/10.3390/app9194050
4. N. Li, S. Liu, Y. Liu, S. Zhao, M. Liu, Neural Speech Synthesis with Transformer Network, in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33. PKP Publishing, Honolulu, Hawaii, pp. 6706–6713 (2019). https://doi.org/10.1609/aaai.v33i01.33016706
5. J. Shen, R. Pang, R.J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R.A. Saurous, Y. Agiomvrgiannakis, Y. Wu, Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions, in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Calgary, AB, Canada, pp. 4779–4783 (2018). https://doi.org/10.1109/icassp.2018.8461368
6. Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, T.-Y. Liu, Fastspeech: Fast, robust and controllable text to speech. Advances in Neural Information Processing Systems **32** (2019)
7. R.C. Guido, Wavelets behind the scenes: Practical aspects, insights, and perspectives. Phys. Rep. **985**, 1–23 (2022). https://doi.org/10.1016/j.physrep.2022.08.001
8. X. Zheng, Y. Tang, J. Zhou, A framework of adaptive multiscale wavelet decomposition for signals on undirected graphs. IEEE Trans. Signal Process. **67**, 1696–1711 (2019). https://doi.org/10.1109/tsp.2019.2896246
9. L. Yang, H. Su, C. Zhong, Z. Meng, H. Luo, X. Li, Y.Y. Tang, Y. Lu, Hyperspectral image classification using wavelet transform-based smooth ordering. Int. J. Wavelets Multiresolut. Inf. Process. **17**, 1950050 (2019). https://doi.org/10.1142/s0219691319500504
10. R.C. Guido, Effectively interpreting discrete wavelet transformed signals [lecture notes]. IEEE Signal Process. Mag. **34**, 89–100 (2017). https://doi.org/10.1109/msp.2017.2672759
11. R.C. Guido, Practical and useful tips on discrete wavelet transforms [sp tips & tricks]. IEEE Signal Process. Mag. **32**, 162–166 (2015). https://doi.org/10.1109/msp.2014.2368586
12. E. Guariglia, Primality, fractality, and image analysis. Entropy **21**, 304 (2019). https://doi.org/10.3390/e21030304
13. E. Guariglia, S. Silvestrov, Fractional-wavelet analysis of positive definite distributions and wavelets on D′(C), in *Engineering Mathematics II. Springer Proceedings in Mathematics & Statistics*, vol. 179, ed. by S. Silvestrov, M. Rančić (Springer, Cham, 2016), pp.337–353. https://doi.org/10.1007/978-3-319-42105-6_16
14. E. Guariglia, R.C. Guido, Chebyshev wavelet analysis. J. Funct. Spaces **2022**, 5542054 (2022). https://doi.org/10.1155/2022/5542054

15. S.G. Mallat, A theory for multiresolution signal decomposition: the wavelet representation. IEEE Trans. Pattern Anal. Mach. Intell. **11**, 674–693 (1989). https://doi.org/10.1109/34.192463
16. A. Grossmann, R. Kronland-Martinet, J. Morlet, *Reading and Understanding Continuous Wavelet Transforms. Wavelets. Inverse Problems and Theoretical Imaging* (Springer, Berlin, Heidelberg, 1989), pp.2–20. https://doi.org/10.1007/978-3-642-97177-8_1
17. R. Caruana, Multitask learning. Mach. Learn. **28**, 41–75 (1997). https://doi.org/10.1023/a:1007379606734
18. N. Adiga, S.R.M. Prasanna, Acoustic features modelling for statistical parametric speech synthesis: a review. IETE Tech. Rev. **36**, 130–149 (2019). https://doi.org/10.1080/02564602.2018.1432422
19. Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, T.-Y. Liu, Fastspeech 2: Fast and high-quality end-to-end text to speech. arXiv:2006.04558 [eess.AS] (2022) https://doi.org/10.48550/arXiv.2006.04558
20. Y. Chisaki, H. Nakashima, S. Shiroshita, T. Usagawa, M. Ebata, A pitch detection method based on continuous wavelet transform for harmonic signal. Acoust. Sci. Technol. **24**, 7–16 (2003). https://doi.org/10.1250/ast.24.7
21. S. Kadambe, G.F. Boudreaux-Bartels, Application of the wavelet transform for pitch detection of speech signals. IEEE Trans. Inf. Theory **38**, 917–924 (1992). https://doi.org/10.1109/18.119752
22. A. Mehrish, N. Majumder, R. Bhardwaj, R. Mihalcea, S. Poria, A review of deep learning techniques for speech processing. arXiv:2305.00359 [eess.AS] (2023) https://doi.org/10.48550/arXiv.2305.00359
23. K. O'Shea, R. Nash, An introduction to convolutional neural networks. arXiv:1511.08458 [cs.NE] (2015) https://doi.org/10.48550/arXiv.1511.08458
24. A. Sherstinsky, Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. Phys. D Nonlinear Phenom. **404**, 132306 (2020). https://doi.org/10.1016/j.physd.2019.132306
25. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need. Advances in Neural Information Processing Systems **30** (2017)
26. S.O. Arık, M. Chrzanowski, A.Coates, G. Diamos, A. Gibiansky, Y.Kang, X. Li, J. Miller, A. Ng, J. Raiman, S. Sengupta, M. Shoeybi, Deep Voice: Real-time Neural Text-to-Speech, in *Proceedings of the 34th International Conference on Machine Learning*, vol. 70. PMLR, Sydney, Australia, pp. 195–204 (2017). https://proceedings.mlr.press/v70/arik17a.html
27. A. Gibiansky, S. Arik, G. Diamos, J. Miller, K. Peng, W. Ping, J. Raiman, Y. Zhou, Deep voice 2: Multi-speaker neural text-to-speech. Advances in Neural Information Processing Systems **30** (2017)
28. W. Ping, K. Peng, A. Gibiansky, S.O. Arik, A. Kannan, S. Narang, J. Raiman, J. Miller, Deep voice 3: Scaling text-to-speech with convolutional sequence learning. arXiv:1710.07654 [cs.SD] (2018) https://doi.org/10.48550/arXiv.1710.07654
29. S. Ouelha, A. Aïssa-El-Bey, B. Boashash, An improved time-frequency noise reduction method using a psychoacoustic mel model. Digit. Signal Process. **79**, 199–212 (2018). https://doi.org/10.1016/j.dsp.2018.04.005
30. J. Xiao, J. Liu, D. Li, L. Zhao, Q. Wang, Speech Intelligibility Enhancement By Non-Parallel Speech Style Conversion Using CWT and iMetricGAN Based CycleGAN, in *MultiMedia Modeling. MMM 2022. Lecture Notes in Computer Science*, vol. 13141. Springer, Cham, pp. 544–556 (2022). https://doi.org/10.1007/978-3-030-98358-1_43
31. Y. Gu, Y. Kang, Multi-task wavenet: A multi-task generative model for statistical parametric speech synthesis without fundamental frequency conditions. arXiv:1806.08619 [eess.AS] (2018) https://doi.org/10.48550/arXiv.1806.08619
32. Z. Huang, J. Li, S.M. Siniscalchi, I.-F. Chen, J. Wu, C.-H. Lee, Rapid adaptation for deep neural networks through multi-task learning, in *Sixteenth Annual Conference of the International Speech Communication Association*. INTERSPEECH, ISCA, Dresden, Germany, pp. 3625–3629 (2015). https://doi.org/10.21437/interspeech.2015-719
33. Z. Wu, C. Valentini-Botinhao, O. Watts, S. King, Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis, in 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, South Brisbane, QLD, Australia, pp. 4460–4464 (2015). https://doi.org/10.1109/icassp.2015.7178814
34. J. Chen, L. Ye, Z. Ming, Mass: Multi-task anthropomorphic speech synthesis framework. Comput. Speech Lang. **70**, 101243 (2021). https://doi.org/10.1016/j.csl.2021.101243
35. J.-T. Huang, J. Li, D. Yu, L. Deng, Y. Gong, Cross-language Knowledge Transfer Using Multilingual Deep Neural Network with Shared Hidden Layers, in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, Vancouver, BC, Canada, pp. 7304–7308 (2013). https://doi.org/10.1109/icassp.2013.6639081
36. C.J. Peng, Y.L. Shen, Y.J. Chan, C. Yu, Y. Tsao, T.S. Chi, Perceptual Characteristics Based Multi-objective Model for Speech Enhancement, in Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, vol. 2022-September, Incheon, Korea. ISCA, pp. 211–215 (2022). https://doi.org/10.21437/interspeech.2022-11197
37. J. Lee, S. Han, H. Cho, W. Jung, PHASEAUG: a differentiable augmentation for speech synthesis to simulate one-to-many mapping, in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Rhodes Island, Greece, pp. 1–5 (2023). https://doi.org/10.1109/ICASSP49357.2023.10096374. https://ieeexplore.ieee.org/abstract/document/10096374
38. G. Pironkov, S.U. Wood, S. Dupont, Hybrid-task learning for robust automatic speech recognition. Comput. Speech Lang. **64**, 101103 (2020). https://doi.org/10.1016/j.csl.2020.101103
39. S. Imai, K. Sumita, C. Furuichi, Mel log spectrum approximation (mlsa) filter for speech synthesis. Electron. Commun. Japan (Part I Commun.) **66**, 10–18 (1983). https://doi.org/10.1002/ecja.4400660203
40. P. Stoica, R.L. Moses, *Spectral Analysis of Signals*, vol. 452 (Pearson Prentice Hall, Upper Saddle Riverr, 2005)
41. I. Daubechies, The wavelet transform, time-frequency localization and signal analysis. IEEE Trans. Inf. Theory **36**, 961–1005 (1990). https://doi.org/10.1109/18.57199
42. A. Rai, S.H. Upadhyay, A review on signal processing techniques utilized in the fault diagnosis of rolling element bearings. Tribol. Int. **96**, 289–306 (2016). https://doi.org/10.1016/j.triboint.2015.12.037
43. H. Soltau, G. Saon, T.N. Sainath, Joint Training of Convolutional and Non-convolutional Neural Networks, in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Florence, Italy, pp. 5572–5576 (2014). https://doi.org/10.1109/icassp.2014.6854669

44. M. Farge, Wavelet transforms and their applications to turbulence. Annu. Rev. Fluid Mech. **24**, 395–458 (1992). https://doi.org/10.1146/annurev.fl.24.010192.002143
45. S. Mallat, W.L. Hwang, Singularity detection and processing with wavelets. IEEE Trans. Inf. Theory **38**, 617–643 (1992). https://doi.org/10.1109/18.119727
46. M.S. Ribeiro, O. Watts, J. Yamagishi, R.A.J. Clark, Wavelet-based Decomposition of F0 as a Secondary Task for DNN-based Speech Synthesis with Multi-task Learning, in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Shanghai, China, pp. 5525–5529 (2016). https://doi.org/10.1109/ICASSP.2016.7472734. https://ieeexplore.ieee.org/abstract/document/7472734
47. I. Daubechies, *Ten Lectures on Wavelets* (Society For Industrial And Applied Mathematics, Philadelphia, 1992)
48. L. Cohen, *Time-Frequency Analysis* (Prentice Hall PTR, Upper Saddle River, 1995)
49. S. Qin, Z. Ji, Multi-resolution time-frequency analysis for detection of rhythms of EEG signals, in *3rd IEEE Signal Processing Education Workshop. 2004 IEEE 11th Digital Signal Processing Workshop, 2004*. IEEE, Taos Ski Valley, NM, USA, pp. 338–341 (2004). https://doi.org/10.1109/DSPWS.2004.1437971. https://ieeexplore.ieee.org/abstract/document/1437971
50. K. Ito, L. Johnson, The LJ Speech Dataset (2017). https://keithito.com/LJ-Speech-Dataset/
51. G.H. Golub, C. Reinsch, Singular value decomposition and least squares solutions. Numer. Math. **14**, 403–420 (1970). https://doi.org/10.1007/bf02163027
52. S. Ioffe, C. Szegedy, Batch normalization: accelerating deep network training by reducing internal covariate shift, in *Proceedings of the 32nd International Conference on Machine Learning*, vol. 37. PMLR, Lille, France, pp. 448–456 (2015). http://proceedings.mlr.press/v37/ioffe15.html
53. M. Schuster, K.K. Paliwal, Bidirectional recurrent neural networks. IEEE Trans. Signal Process. **45**, 2673–2681 (1997). https://doi.org/10.1109/78.650093
54. S. Hochreiter, J. Schmidhuber, Long short-term memory. Neural Comput. **9**, 1735–1780 (1997). https://doi.org/10.1162/neco.1997.9.8.1735

Publisher's Note