SAMPLERNN-BASED NEURAL VOVODER FOR STATISTICAL PARAMETRIC SPEECH SYNTHESIS

Yang Ai, Hong-Chuan Wu, Zhen-Hua Ling

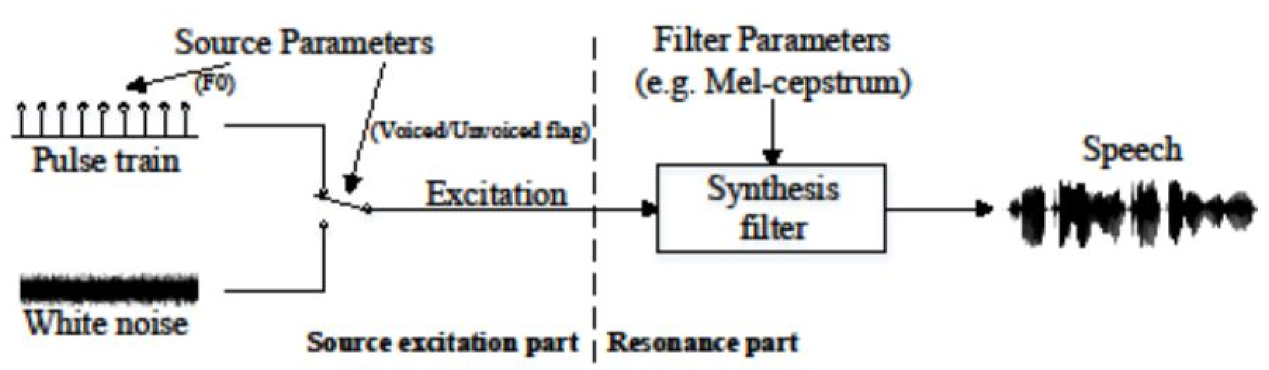
National Engineering Laboratory for Speech and Language Information Processing, University of Science and Technology of China, Hefei, P.R.China

Abstract

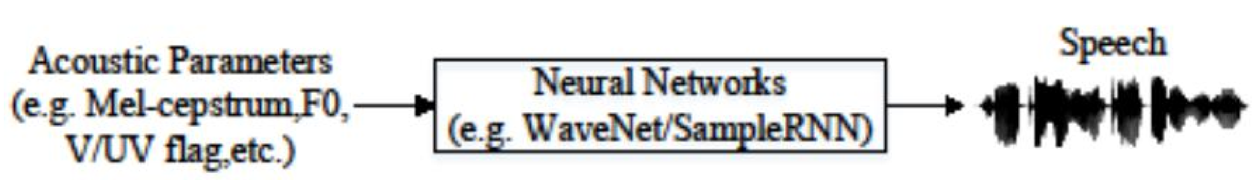
- This paper presents a SampleRNN-based neural vocoder for SPSS.
- The model is composed of a hierarchical structure of GRU layers and feed-forward layers.
- The model can capture long-span dependencies between acoustic features and waveform sequences.
- The waveform samples are generated in an autoregressive manner.
- Objective and subjective performance: the vocoder outperform WaveNet-based neural vocoder and STRAIGHT.

Proposed Method

- **Comparsion of neural vocoder and conventional vocoder**



(a) Conventional vocoder



(b) Neural vocoder

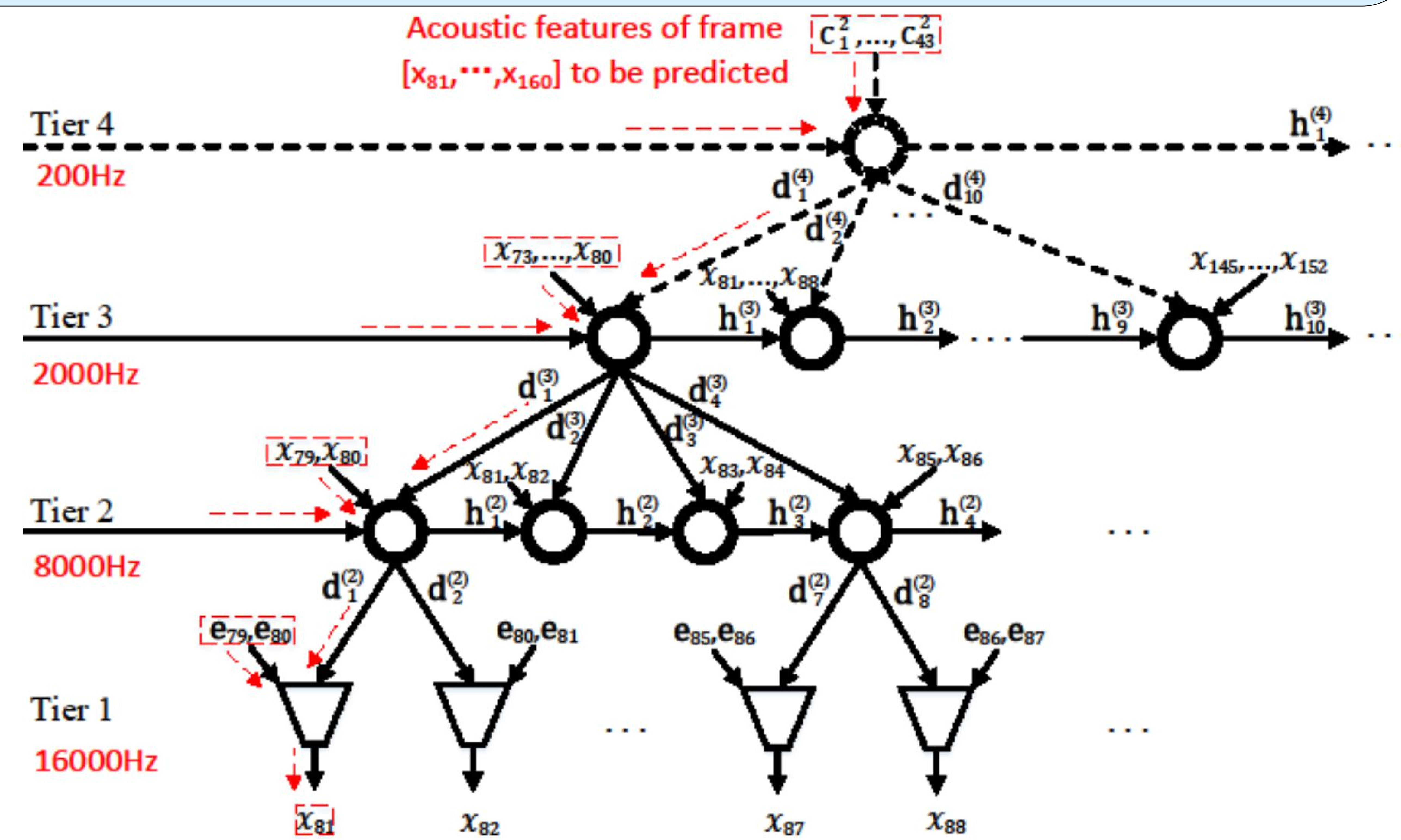
✓ Conventional vocoder: based on the source-filter model. The vocoder (e.g. STRAIGHT) losts the spectral details and phase information and ignores the nonlinear effects in practical speech production.
✓ Neural vocoder: convert acoustic parameters into speech by a designed neural network (e.g. WaveNet and SampleRNN) directly. The neural vocoder can overcome the deficiencies of conventional vocoder.

- **Basic unconditional SampleRNN**

✓ Solid line in figure
✓ A waveform generator composed of a hierarchical structure of GRU layers and FF layers in an autoregressive manner
✓ Generate one sample conditioned on its previous samples

- **SampleRNN-based neural vocoder**

✓ Figure: conditional SampleRNN model
✓ Dotted lines represent the conditional tier added on the top of basic unconditional SampleRNN
✓ The input of conditional tier is acoustic features of one frame of samples to be predicted
✓ Train to Minimize the cross-entorpy
✓ Generate one sample conditioned on its previous samples and its corresponding acoustic features



Experiments

- **Conditions**

Database	Chinese corpus with 1000 utterances from a female speaker and English corpus with 1000 utterances from a male speake. training/validation/test set: 800/100/100
Acoustic Features	Composition: 40-order MCCs,1-order power, 1-order F0,and 1-order binary U/V flag. Type: natural features (R) and predicted features (P).
Systems	STRAIGHT, WaveNet, SampleRNN

- **Comparison of classfication accuracy (ACC) and cross entropy (CE) on test set**

	Chinese famle		English male	
	WaveNet	SampleRNN	WaveNet	SampleRNN
ACC(%)	19.77	20.59	14.16	14.51
CE	2.7427	2.6983	3.2304	3.1570

✓ SampleRNN > WaveNet

- **Comparison of distortion on the test set of the Chinese corpus**

	STRAIGHT	WaveNet	SampleRNN
SNR(dB)	2.4994	4.7093	5.1987
MCD(dB)	1.5744	1.6919	1.4950
F0-RMSE (cent)	20.6821	14.9475	11.4926
V/UV error (%)	2.9172	3.5552	3.1725

✓ SNR: distortion in time domain
✓ MCD: distortion in mel-cepstrum
✓ F0-RMSE and V/UV error: distortion in F0
✓ SampleRNN > WaveNet> STRAIGHT
✓ From SNR, neural vocoders can recover pahse information more accurately.

✓ Note: Results in English corpus shown in paper

- **Average preference scores (%) on speech quality using the Chinese corpus**

	STRAIGHT	WaveNet	SampleRNN	N/P
R	10.55	--	55.05	34.40
	--	9.17	37.16	53.67
P	9.13	--	54.80	36.07
	--	10.18	38.89	50.93

✓ N/P: no preference
✓ SampleRNN > STRAIGHT
✓ SampleRNN > WaveNet
✓ p-values of a t-test are all less than 0.001
✓ For predicted features as input, SampleRNN-based vocoder has better performance.
✓ Time consumed for generating one second speech was 91.89s for the SampleRNN-based neural vocoder