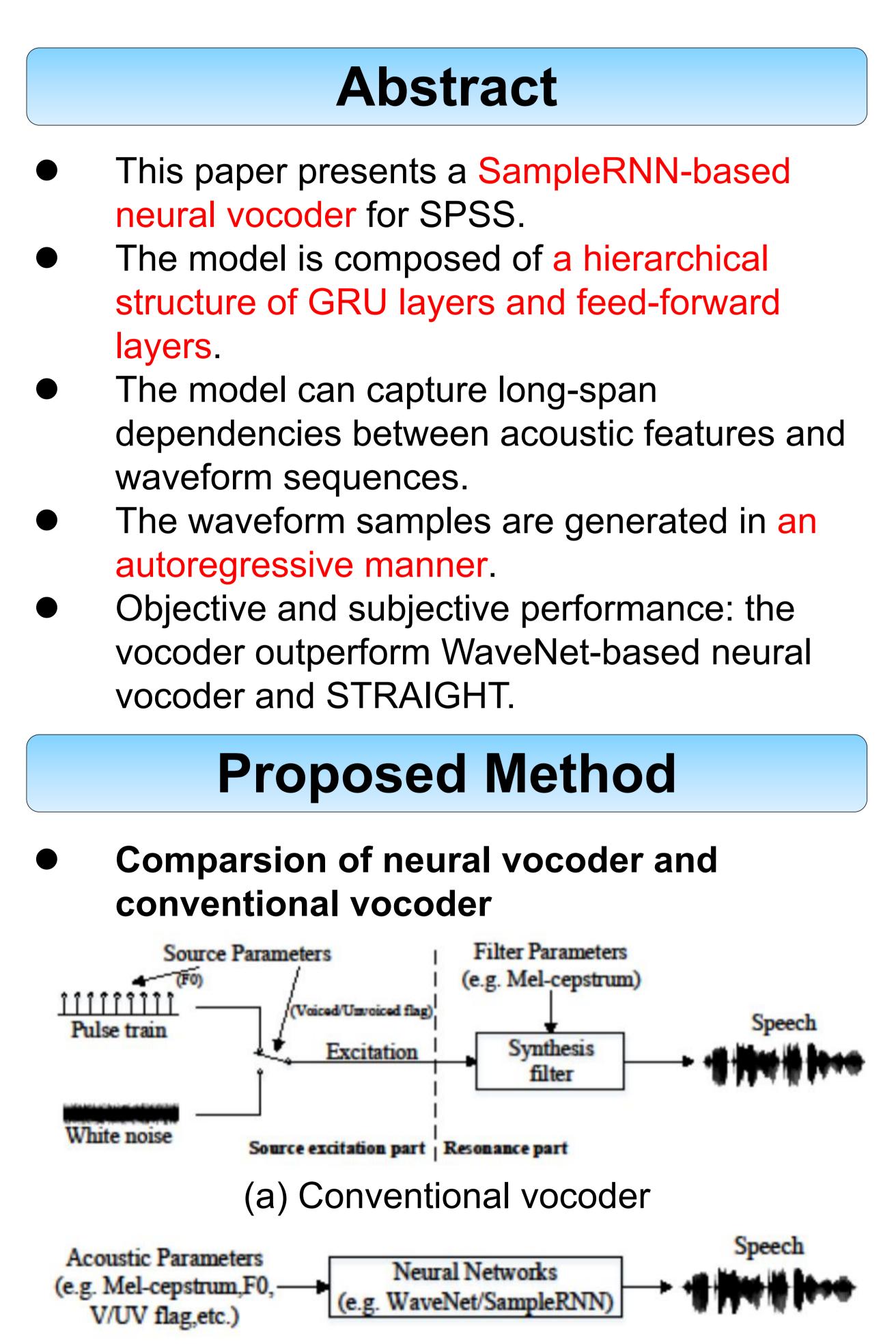
SAMPLERNN-BASED NEURAL VOVODER FOR STATISTICAL PARAMETRIC SPEECH SYNTHESIS Yang Ai, Hong-Chuan Wu, Zhen-Hua Ling

University of Science and Technology of China, Hefei, P.R.China



(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.

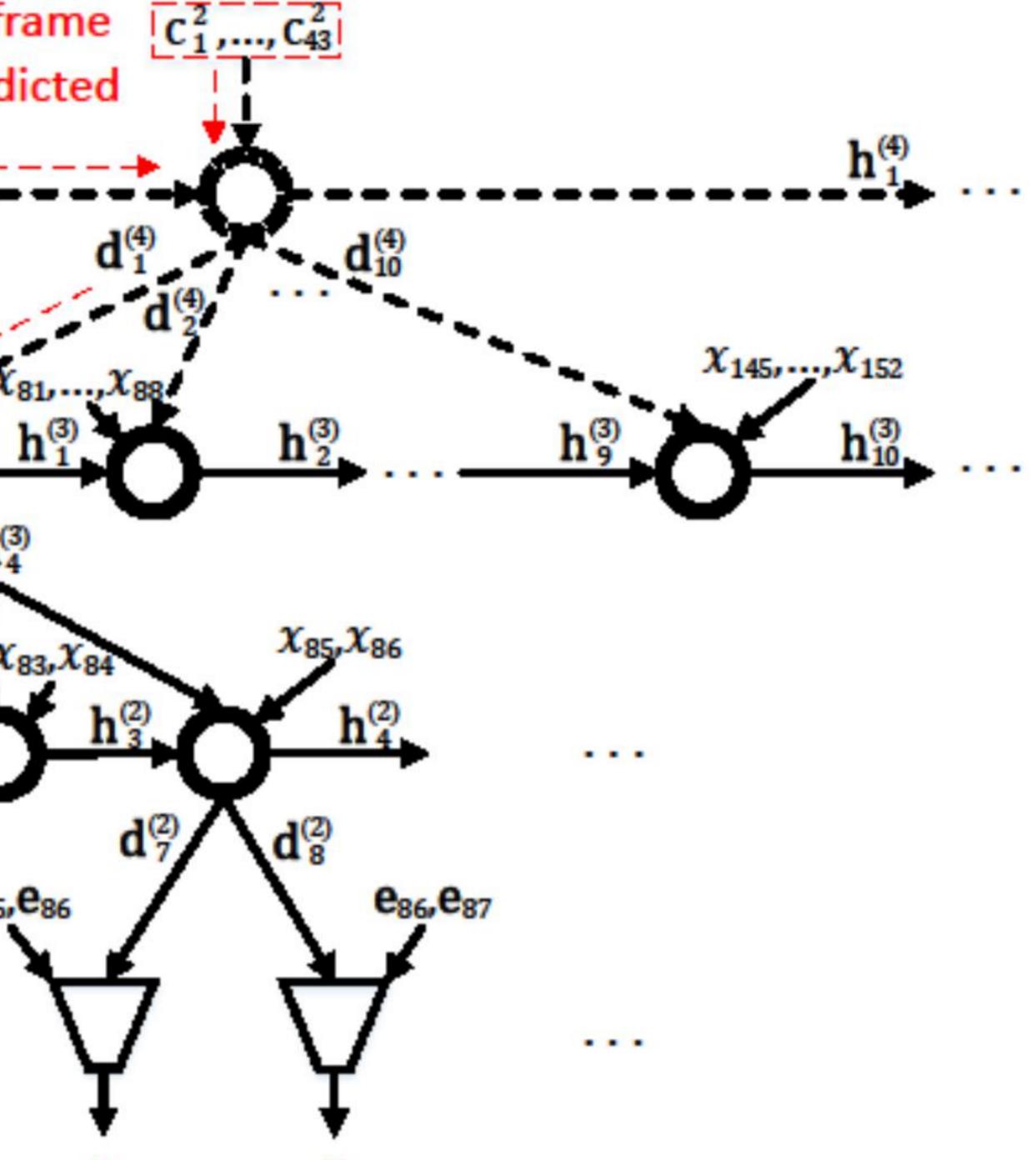
National Engineering Laboratory for Speech and Language Information Processing,

	ia, iicici,	Г.Л.С	ΙΠΠα										
							Acoustic	features of fra	ame [c12,	, C ₄₃			
• Bas	sic uncondi ⁻	tional Sa	ampleRNN					60] to be predi					
 ✓ 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 					Tier 4 200Hz			,X ₈₀	d ⁽⁴⁾ d ⁽⁴⁾ d ⁽⁴⁾ 2	d ⁽⁴⁾	•••••	h ⁽⁴⁾	
	ous samples	•			Tier 3				(3)	h ⁽³⁾	h ⁽³⁾	h ⁽³⁾	
 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 					2000Hz	X 79, X 80	X ₈₁ X	$d_{1}^{(3)}$ $d_{4}^{(3)}$ $d_{2}^{(3)}$ $d_{3}^{(3)}$ $d_{3}^{(3)}$ χ_{2}	x ₁ 3, χ ₈₄	x ₈₅ , x ₈₆			
					Tier 2 8000Hz Tier 1	d ⁽²⁾ e ₇₉ ,e ₈₀	h ⁽²⁾ d ⁽²⁾ e ₈₀	h ⁽²⁾ ,e ₈₁ e ₈₅ ,e	h ⁽²⁾ d ⁽²⁾ e ₈₆	h ⁽²⁾ d ⁽²⁾ e ₈₆ ,e ₈₇			
	Exr	oerim	ents		16000Hz	Y	Y		Y	Y			
• Co	- nditions					X81	X82		X87	X88			
Chinese corpus with 1000 utterances from a female speaker and English and corpus with			 Comparison of distortion on the test set of the Chinese corpus Average preference scores (%) on speed quality using the Chinese corpus 								eech		
Database	1000 uttera		n a male spea st set [.] 800/10			STRAIGHT	WaveNet	SampleRNN		STRAIGHT	WaveNet	SampleRNN	N/P
	training/validation/test set: 800/100/100 Composition: 40-order MCCs,1-order power,			SNR(dB)	2.4994	4.7093	5.1987	R	10.55		55.05	34.40	
Acoustie Feature	 c 1-order F0, and 1-order binary U/V flag. s Type: natural features (R) and predicted 				MCD(dB)	1.5744	1.6919	1.4950			9.17	37.16	53.67
Systems	features (P	?) .			F0-RMSE (cent)	20.6821	14.9475	11.4926		9.13		54.80	36.07
			at CompleDN	INT					Р		40.40		
5	SIRAIGHI		et, SampleRN ation accur		V/UV error	2.9172	3.5552	3.1725	P		10.18	38.89	50.93
• Cor	STRAIGHT STRAIGHT Sparison of C) and cros	classfic	ation accur	acy	V/UV error (%) ✓ SNR: distor	tion in time d	lomain	3.1725	 ✓ N/P: ✓ Samp 	no preferenc oleRNN > S1	e RAIGHT	38.89	50.93
• Cor	nparison of C) and cros	classfic s entrop	ation accur	acy est set male	 V/UV error (%) ✓ SNR: distor ✓ MCD: distor 	rtion in time d rtion in mel-c	lomain epstrum		 ✓ N/P: ✓ Samp ✓ Samp 	no preferenc oleRNN > S1 oleRNN > Wa	e FRAIGHT aveNet		
• Cor	nparison of C) and cros	classfic s entrop	ation accur by (CE) on to	acy est set	 V/UV error (%) ✓ SNR: distor ✓ MCD: distor ✓ F0-RMSE a ✓ SampleRNI 	tion in time d rtion in mel-c and V/UV erro V > WaveNe	Iomain epstrum or: distortion	in F0	 ✓ ✓	no preference oleRNN > ST oleRNN > Wa ues of a t-tes redicted feat	RAIGHT RAIGHT aveNet st are all le tures as in	ess than 0.00 put, Sample	1 RNN-
• Cor	nparison of C) and cros	classfic s entrop	ation accur by (CE) on to English	acy est set male SampleR	 V/UV error (%) ✓ SNR: distor ✓ MCD: distor ✓ F0-RMSE a ✓ SampleRNI ✓ From SNR, 	tion in time d rtion in mel-c and V/UV erro V > WaveNe	Iomain epstrum or: distortion t> STRAIGI ders can rec	in F0	 ✓ N/P: ✓ N/P: ✓ Samp ✓ Samp ✓ P-valu ✓ For p based ✓ Time 	no preference oleRNN > ST oleRNN > Wa ues of a t-tes redicted feat d vocoder ha consumed feat	RAIGHT RAIGHT aveNet st are all le tures as in as better p or generat	ess than 0.00	1 RNN-

	a, neiei	, F.N.C	ΠΠα											
							Acoustic	features of fr	ame	C ₁ ² ,	, C ₄₃			
• Bas	sic uncond	itional Sa	ampleRNN					60] to be pred						
	✓ Solid line in figure								4				h ⁽⁴⁾	
			nposed of a		200Hz				-a(4	4)	-1 ⁽⁴⁾			
			RU layers a	nd FF						1 (4)	u 10			
•	Iayers in an autoregressive manner ✓ Generate one sample conditioned on its						X73	3,,X ₈₀	· ·	u 2			x_{145}, \dots, x_{152}	
	previous samples								81,,χ ₈	88	h ⁽³⁾	h(3)	h ⁽³⁾	
• San	SampleRNN-based neural vocoder)—	1 2	119		
✓ Figure	✓ Figure: conditional SampleRNN model				2000Hz				0	_				
	✓ Dotted lines represent the conditional tier added					X79, X80		$d_2^{(3)}$ $d_3^{(3)}$			X85, X86			
	on the top of basic unconditional SampleRNN						X ₈₁ ,X	82 X	83, X 84		1 (2)			
The input of conditional tier is acoustic features of one frame of samples to be predicted					Tier 2		h ^w ₁		h ^w ₃	40	h ₄			
✓ Train	 Train to Minimize the cross-entorpy 					d ⁽²⁾	⊼d [⊘]		d	[®] X	d ⁽²⁾			
		•	ditioned on i			e79, e80	e 8	o,e ₈₁ e ₈₅ ,	e ₈₆	7	e ₈₆ ,e ₈₇	7		
•	•	s and its (correspondir	ng acoustic						_				
TCatur	features								∇	7	∇			
	Experiments					Ţ	¥		۲		Y			
• Cor	nditions					X 81	X82		X 87	,	x ₈₈			
	Chinese corpus with 1000 utterances from a			 Comparison of distortion on the test set of Average preference scores (%) on speech 										
Dotobooc	female sp	female speaker and English and corpus with 1000 utterances from a male speake.			the Chin	ese corpus				qualit	y using the	e Chinese	e corpus	
Database	1000 utter					STRAIGHT	WaveNet	SampleRNN			STRAIGHT	WaveNet	SampleRNN	N/P
	U	 training/validation/test set: 800/100/100 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). 			SNR(dB)	2.4994	4.7093	5.1987			10.55		55.05	34.40
Acoustic					MCD(dB)	1.5744	1.6919	1.4950		R		9.17	37.16	53.67
reatures					F0-RMSE	20.6821	14.9475	11.4926			9.13		54.80	36.07
Systems	STRAIGH	T, WaveNe	et, SampleRN	١N	- (cent)					Ρ		10.18	38.89	50.93
	 Systems STRAIGHT, WaveNet, SampleRNN Comparison of classification accuracy 			V/UV error (%)	2.9172	3.5552	3.1725			_				
(AC	C) and cro	ss entrop	oy (CE) on t	test set		tion in time d	lomoin				o preference IeRNN > S			
	Chinese famle English male			 ✓ SNR: distortion in time domain ✓ MCD: distortion in mel-cepstrum 				\checkmark	•	leRNN > W				
				✓ F0-RMSE and V/UV error: distortion in F0					•			ess than 0.00		
	WaveNet	NaveNet NN WaveNet NN			 ✓ SampleRNN > WaveNet> STRAIGHT ✓ From SNR, neural vocoders can recover pahse ✓ From SNR, neural vocoders can recover pahse ✓ From SNR, neural vocoders can recover pahse 									
ACC(%)	19.77	20.59	14.16	14.51	 From SNR, neural vocoders can recover pahse information more accurately. 							•	ting one sec	
CE	2.7427	2.6983	3.2304	3.1570								U	SampleRNN	
√ Sam	pleRNN >	WaveNet			Note: Resu	Its in English	corpus sho	wn in paper		neural	vocoder			

SampleRNN > WaveNet





- neural vocoder