Statistical speech synthesis:
Neural is not enough
(at least sometimes)

HUAWEI TECHNOLOGIES CO. LTD (Russia)

Mikhail Kudinov

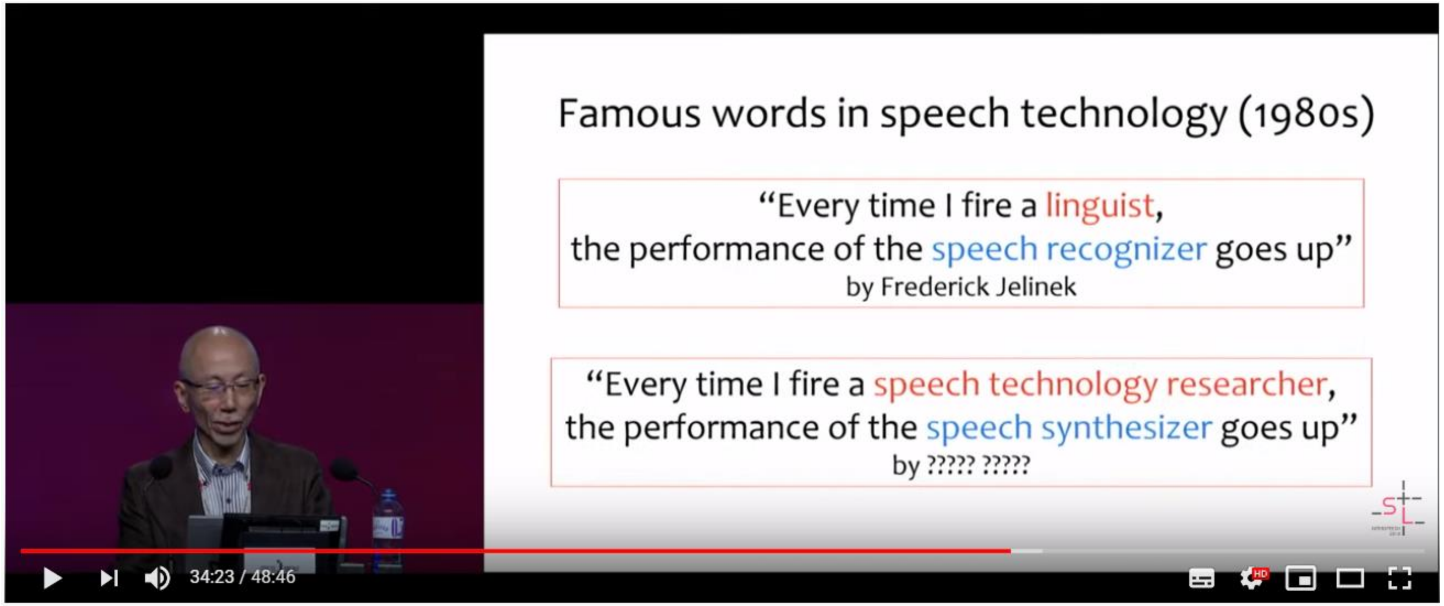
www.huawei.com

HUAWEI

Contents:

1. **What is speech**
2. **Why knowing deep learning is enough for good TTS system**
3. **Why sometimes it might not be true**

Preface: No country for old men

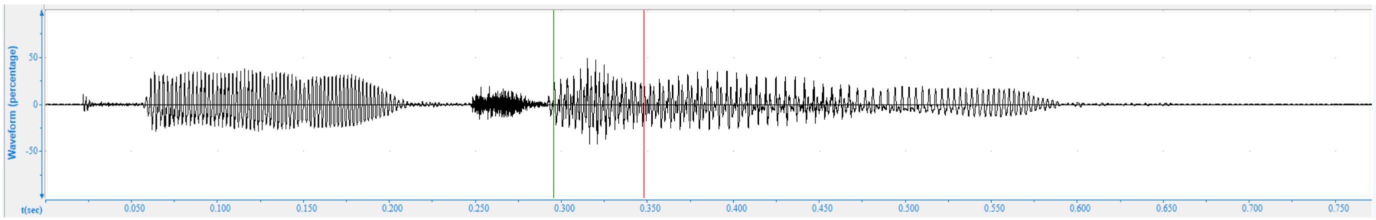


Interspeech 2019 -- Keynote -- Keiichi Tokuda: Statistical approach to speech synthesis

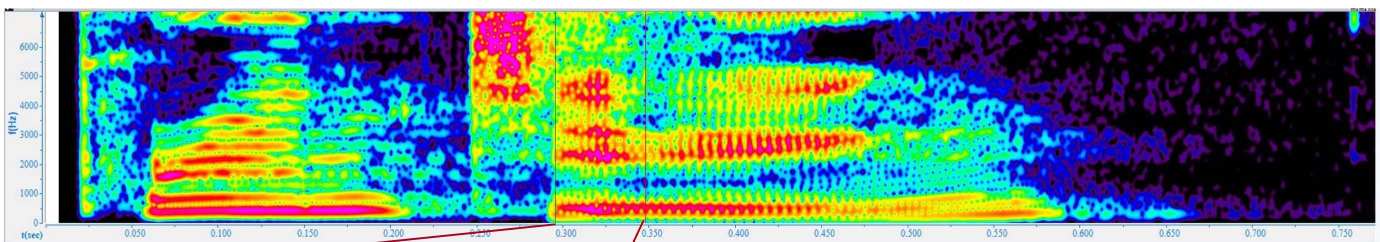
Speech representation and production

WHAT IS SPEECH?

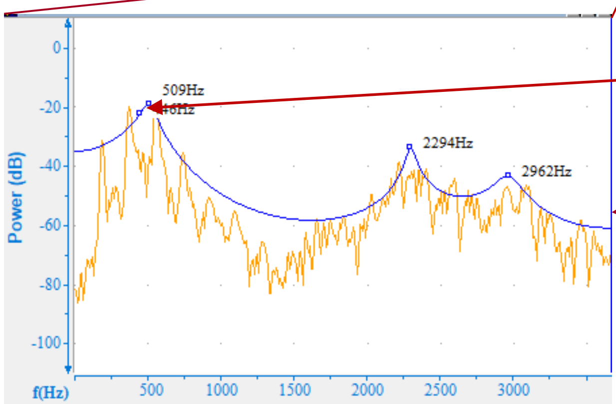
Waveform and Spectrogram



Waveform



Spectrogram

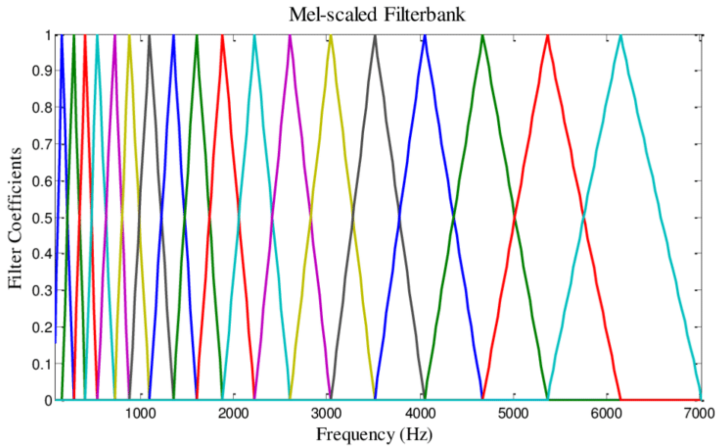


Formant frequencies

Fast Fourier Transform
(10-30 msec segment)

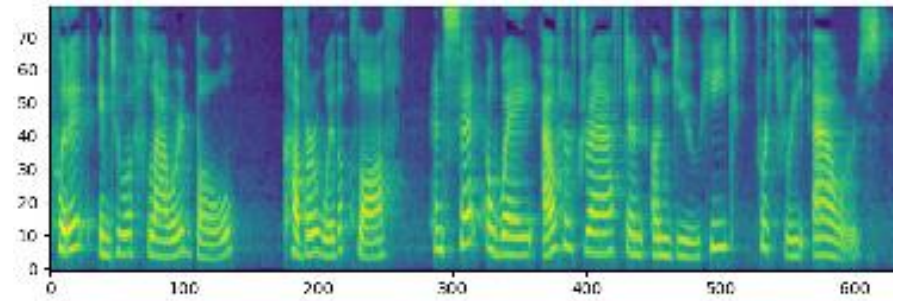
Closer to human perception: Mel-scale

- FFT coefficients are redundant
- Sound perception by the human auditory system is highly non-linear
- Idea #1: compress spectrum into a small number of energies in critical subbands
  - ⇨ Mel-scale spectral coefficients (MFSC)
  - ⇨ Bark-scale spectral coefficients

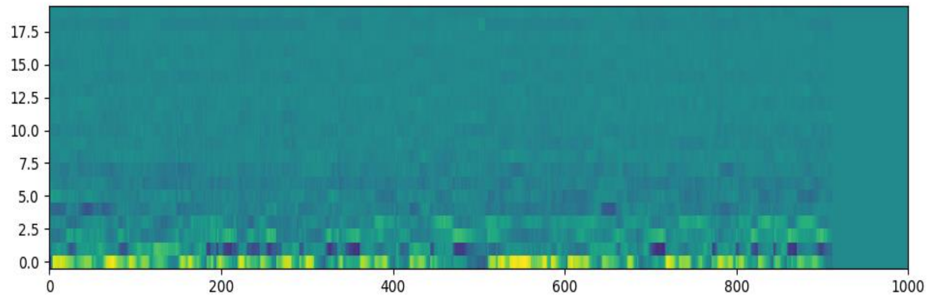


- Idea #2: decorrelate log(MFSC) or log(BFSC) in PCA-like manner (use DCT)
  - ⇨ Mel-frequency cepstral coefficients
  - ⇨ Bark-frequency cepstral coefficients

Acoustic representations: MFSC vs. BFCC



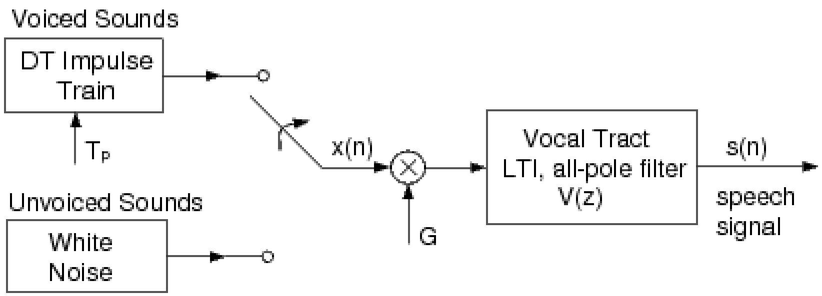
80-band MFSC



18-band Bark-scale cepstrum

Source-Filter model

- Speech signal is a convolution of independent Source (vocal cords) and Filter (vocal tract)
- Linearity assumption: Vocal tract filtering phenomena can be approximately modeled as a linear recursive filter
- Source signal is either an impulse train (voiced sounds) or a white noise (unvoiced sounds)



Vocal tract as a tube

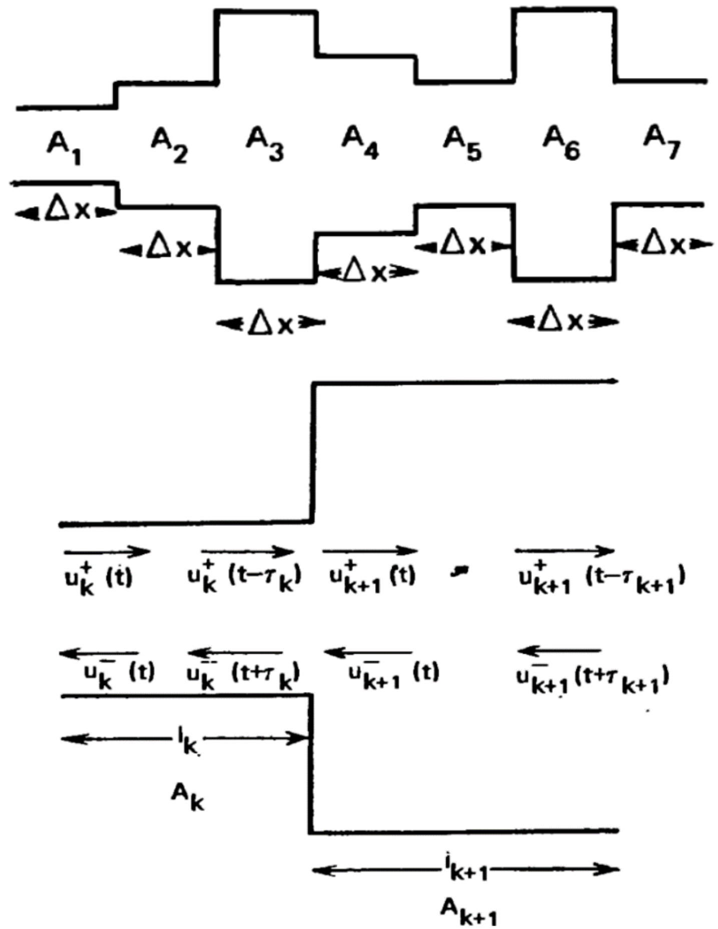
- Let's model the vocal tract as a tube with varying section
- In this case the sound pressure at the end of the last section is a linear combination of multiple running and reflected waves

$$v(t) = a_0\delta(t - N\tau) + \sum_{k=1}^{\infty} a_k\delta(t - N\tau - 2k\tau)$$

- Or in discrete time

$$\mathrm{s}(n) = Gu(n) + \sum_{k=1}^{\infty} a_k s(n-k)$$

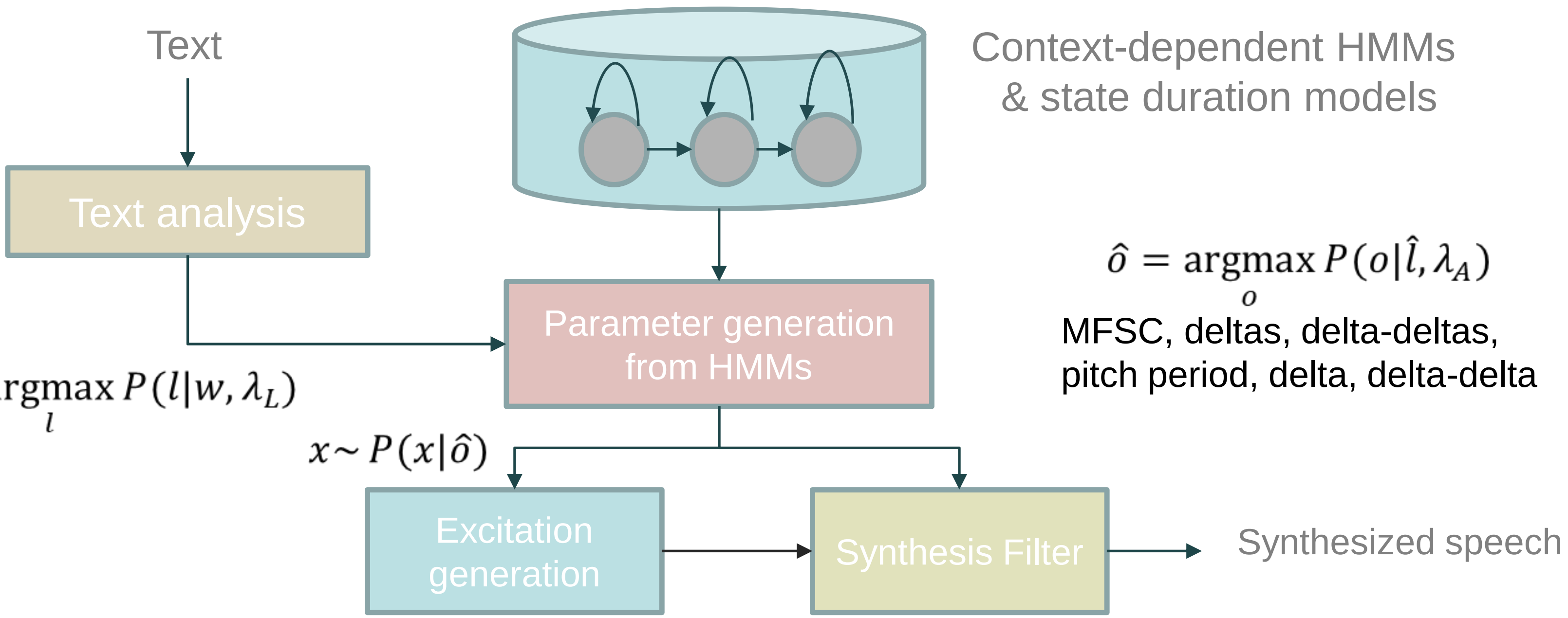
- We can estimate a_k from s(n)
- In this case a_k's are called linear prediction coefficients (LPC)



Unreasonable effectiveness and all the stuff

WHY DEEP LEARNING IS ALL YOU NEED

Statistical speech synthesis before Tacotron era



$$\hat{l} = \operatorname*{argmax}_{l} P(l|w, \lambda_L)$$

$$\hat{o} = \operatorname*{argmax}_{o} P(o|\hat{l}, \lambda_A)$$

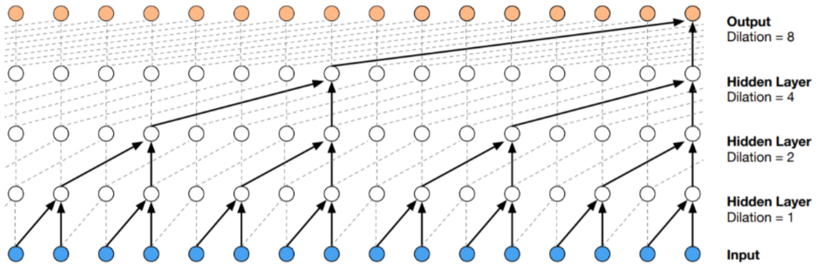
$$x \sim P(x|\hat{o})$$

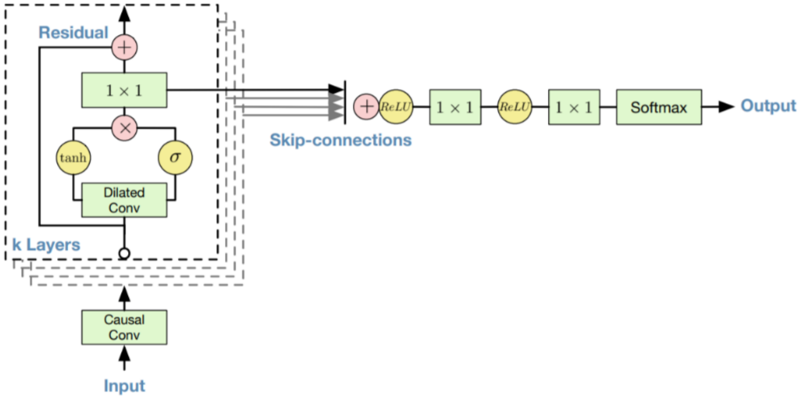
Neural vocoder: WaveNet [Oord17]

- Autoregressive model based
- Main building element: residual block with dilated convolutions
- 30 residual blocks
- Predicts mu-law companded outputs

$$\Rightarrow \frac{\ln(1+\mu|x|)}{1+\ln(1+\mu)}$$

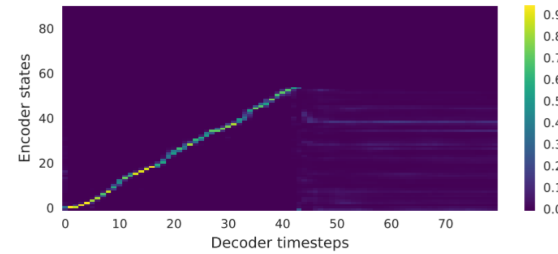
- 16kHz sound
- Best RTF is ~3.0 (Nvidia P100) [Kal18]

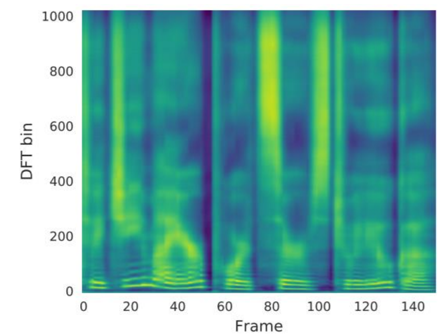




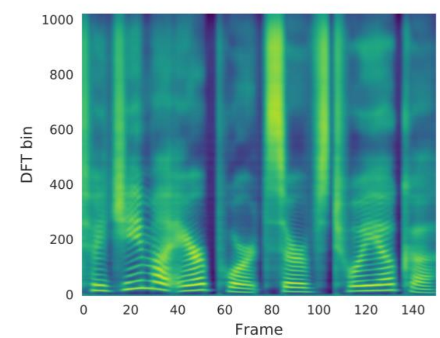
Neural feature-generation: Tacotron/Tacotron2

- Seq2Seq model with attention
  - ⇨ Content-based in Tacotron [Wang17]
  - ⇨ Location-sensitive attention in Tacoton2 [Shen18]
- Predicts 80-band MFSC
  - ⇨ Griffin-Lim vocoder in Tacotron
  - ⇨ WaveNet in Tacotron2
- Predicting N frames per inference step
  - ⇨ N = 2-3 in Tacotron
  - ⇨ N = 1 in Tacoton2 (as reported in [Shen18])
- Makes use of so-called PostNet increasing frequency resolution
- Tacotron2: separate stop-token prediction





(a) Without post-processing net

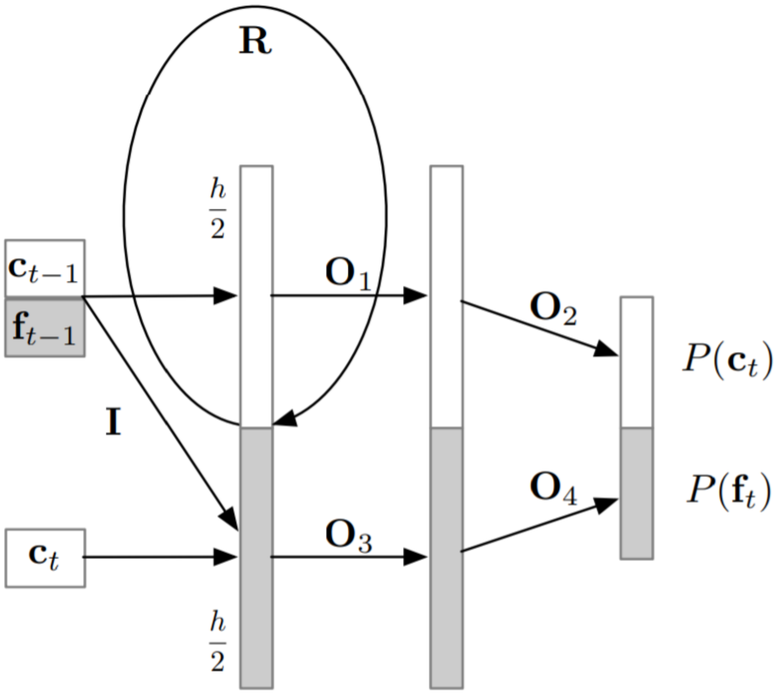


Parallel WaveNet, ClariNet (and WaveGlow)

- Idea #1: use trained autoregressive model as a reference distribution (teacher) to trained a non-autoregressive vocoder (based on IAF) via KL-divergence minimization
- Parallel WaveNet [Oord&Li17]
  - ⇨ MoL
  - ⇨ Power loss, Style loss, Contrastive loss
- ClariNet [Ping19]
  - ⇨ Normal distribution
  - ⇨ STFT loss (almost like power loss)
  - ⇨ Tricks with KL divergence regularization
- WaveGlow [Pren18] uses Glow network and is trained from scratch

Recurrent Vocoders: WaveRNN

- Recurrent neural network conditioned on linguistic features (so, like original WaveNet it's not actually a vocoder) [Kal18]
- Quality comparable to WaveNet but RTF=0.25 on GPU at 24kHz
- 16-bit prediction via 2-step prediction: coarse (first 8-bit) and fine (second 8-bit) parts
- Block sparsification



On desperate living without GPU (and memory)

WHY DEEP LEARNING IS NOT ALL YOU NEED

Going mobile

- Memory consumption
- CPU
- ROM
- Tacotron is considerably fast
- Bottleneck: vocoder
- LPCNet to the rescue!

LPCNet: LPC estimation

- We can estimate a_k by minimizing energy of the residual

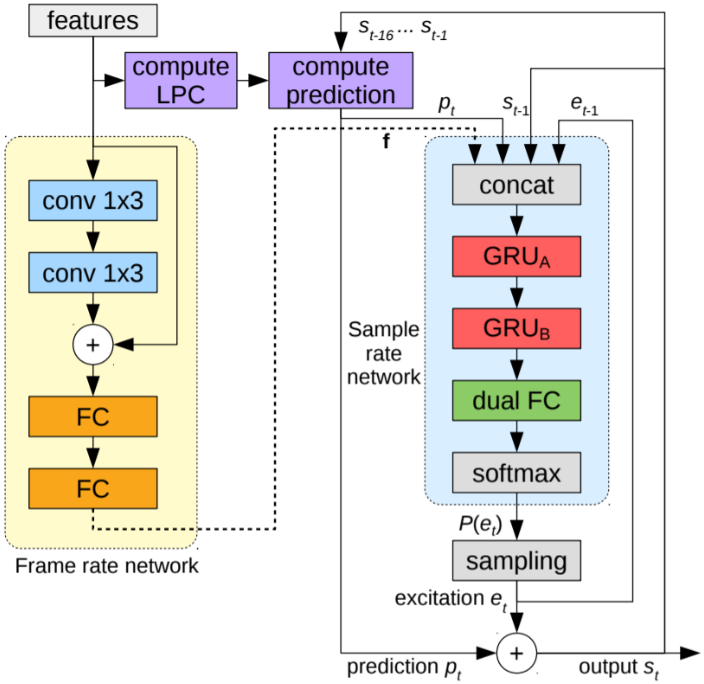
$$E = \sum_n e^2(n) = \sum_n (s(n) - \tilde{s}(n))^2 = \sum_n (s(n) - \sum_{k=1}^{p} a_k s(n-k))^2$$

⇨ which is equivalent to MSE minimization in the analysis window

- LPCs can be found in both time and frequency domain i.e. from spectrum
- If we know e(n) and a_k for the speech frame we can restore s(n)
- Small module of e(n) is a resource of compression in speech codecs

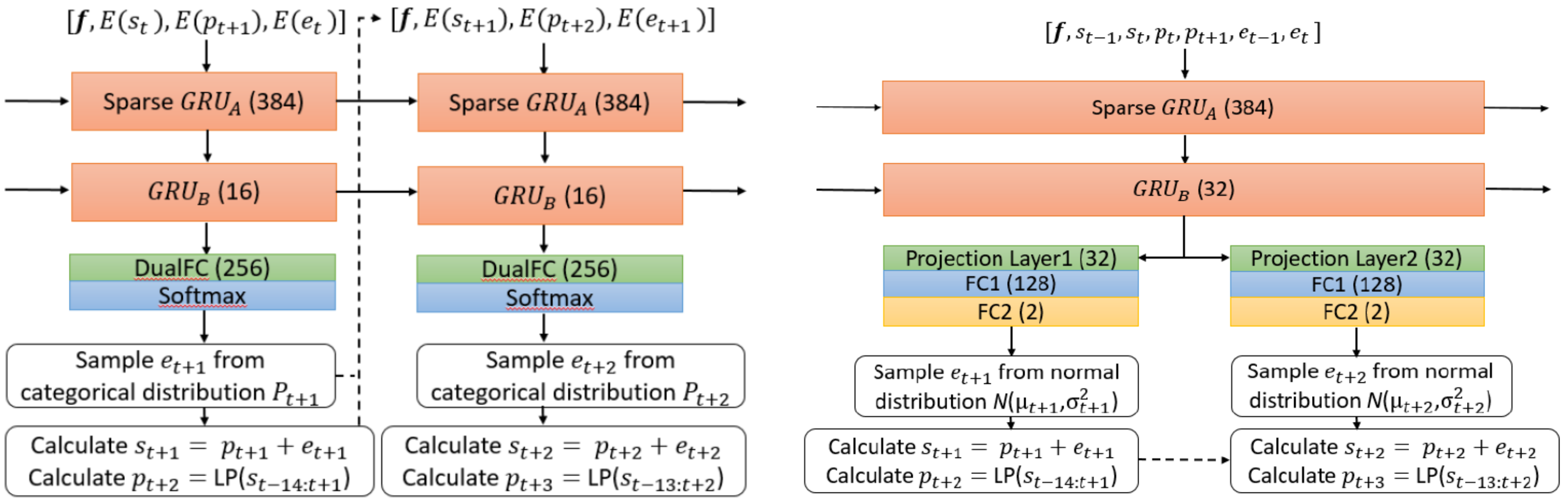
LPCNet

- Compute LPCs from Bark-scale cepstral coefficients
- Use frame-rate network to compute feature vector **f**
- Feed **f** to a sample rate recurrent network predicting excitation signal e(n)
- Output signal is computed as if it was restored from true LPCs and excitation signal
- Operates on 8kHz, 8-bit mu-law
- RTF: ~0.2 on CPU



Val[19]

Multisample Gaussian LPCNet



- Idea #1: predict n>1 samples per step
  - ⇨ Speedup up to 50%
- Idea #2: predict parameters of Normal distribution
  - ⇨ 16-bit sound
  - ⇨ ~50% reduction in number of parameters

Summary

- Tacotron and WaveNet have revolutionized server-based TTS systems
- State-of-art server-based TTS systems are based on Tacotron-like architectures and parallel vocoders (WaveGlow, ClariNet) which require powerful GPU
- CPU-based and mobile services still require a bit of classical DSP techniques to have solid real-time guarantees
- Latest research is focused on emotional multi-speaker and sample-efficient TTS solutions

References

- [Oord17] A. van den Oord et al. WAVENET: A GENERATIVE MODEL FOR RAW AUDIO
- [Kal18] N. Kalchbrenner et al. EFFICIENT NEURAL AUDIO SYNTHESIS
- [Wang17] Y. Wang et al. TACOTRON: TOWARDS END-TO-END SPEECH SYNTHESIS
- [Shen18] J. Shen et al. NATURAL TTS SYNTHESIS BY CONDITIONING WAVENET ON MEL SPECTROGRAM PREDICTIONS
- [Oord&Li17] A. van den Oord et al. PARALLEL WAVENET: FAST HIGH-FIDELITY SPEECH SYNTHESIS
- [Ping18] W.Ping et al. CLARINET: PARALLEL WAVE GENERATION IN END-TO-END TEXT-TO-SPEECH
- [Pren19] R.Prenger et al. WAVEGLOW: A FLOW-BASED GENERATIVE NETWORK FOR SPEECH SYNTHESIS
- [Val19] J.-M. Valin, J. Skoglund LPCNET: IMPROVING NEURAL SYNTHESIS THROUGH LINEAR PREDICTION

Thank You

www.huawei.com

Speech synthesis approaches

- Rule-based, formant synthesis
  - ⇨ Phonetic units are generated according to hand-crafted rules
- Corpus-based, concatenative synthesis
  - ⇨ Concatenate speech units from a database
    - – Diphone synthesis
    - – Unit selection synthesis
- Corpus-based statistical synthesis
  - ⇨ Feature-generation+vocoder
    - – HMM
    - – DNN
  - ⇨ E2E systems

Statistical speech synthesis

- Given a training database consisting of pairs of speech waveforms $\boldsymbol{X}$ and texts $\boldsymbol{W}$ estimate a statistical model parameters λ for generating speech waveform $\mathbf{x}$ for text $\boldsymbol{w} \notin \boldsymbol{W}$: $x \sim P(\boldsymbol{x}|\boldsymbol{w}, \lambda)$
- Usually $P(\boldsymbol{x}|\boldsymbol{w}, \lambda)$ is decomposed into submodules
  - ⇨ $P(\boldsymbol{x}|\boldsymbol{w}, \lambda) = p(\boldsymbol{x}|\boldsymbol{f})\, p(\boldsymbol{f}|\boldsymbol{l}, \lambda_A)\, p(\boldsymbol{l}|\boldsymbol{w}, \lambda_L)$
  - ⇨ **f**: parameteric representation of speech waveform x
  - ⇨ **l**: linguistic feature
  - ⇨ $\lambda = \{\lambda_A, \lambda_L\}$: generative model parameter
    - λ_A : acoustic model parameter
    - λ_L: linguistic model parameter

Tacotron2 architecture layout [Shen18]