Deep Learning in Speech Synthesis

Heiga Zen
Google
August 31st, 2013

Outline

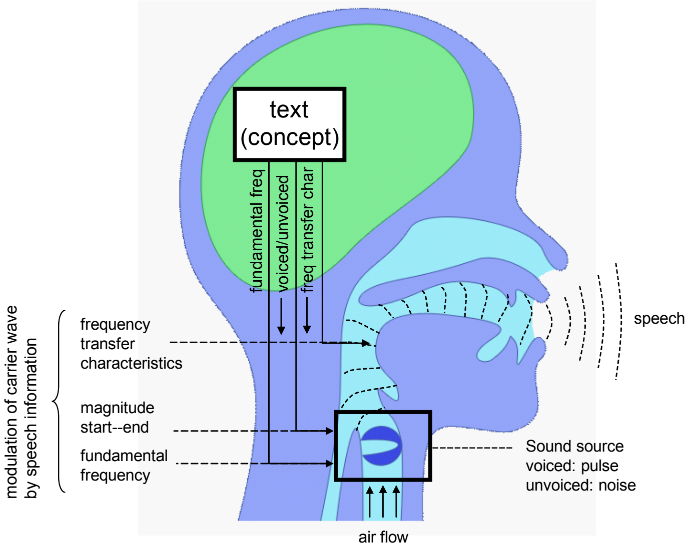
Text-to-speech as sequence-to-sequence mapping

- **Automatic speech recognition (ASR)**
  Speech (continuous time series) → Text (discrete symbol sequence)

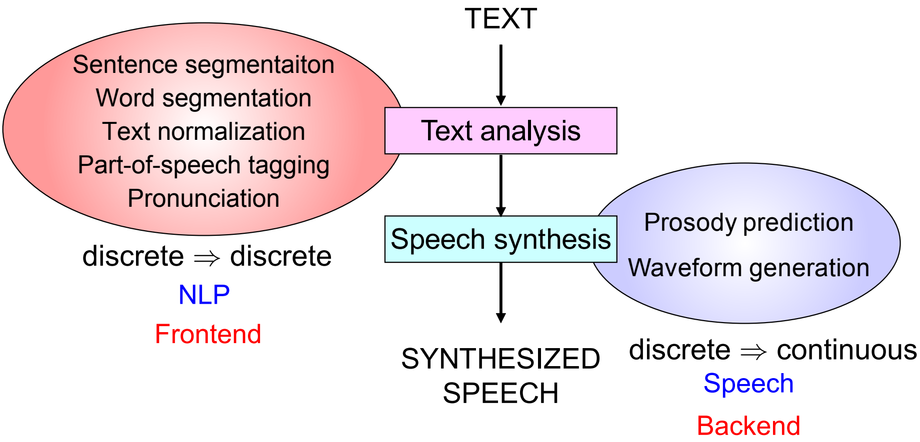
- **Machine translation (MT)**
  Text (discrete symbol sequence) → Text (discrete symbol sequence)

- **Text-to-speech synthesis (TTS)**
  Text (discrete symbol sequence) → Speech (continuous time series)

Speech production process

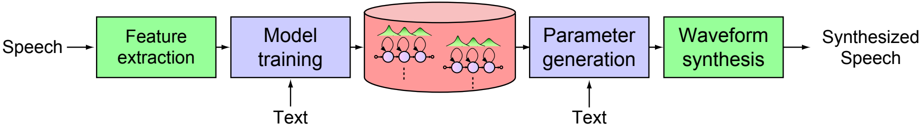


Typical flow of TTS system



This talk focuses on backend

Statistical parametric speech synthesis (SPSS) [2]



- Large data + automatic training
  $\rightarrow$ Automatic voice building
- Parametric representation of speech
  $\rightarrow$ Flexible to change its voice characteristics

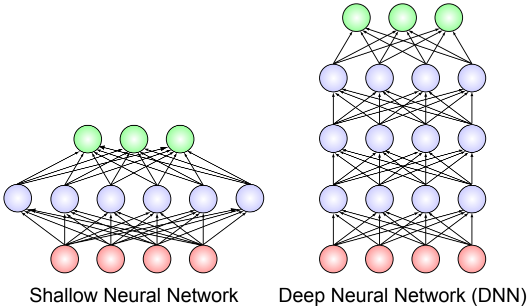
Hidden Markov model (HMM) as its acoustic model
$\rightarrow$ HMM-based speech synthesis system (HTS) [1]

Characteristics of SPSS

- **Advantages**
  - Flexibility to change voice characteristics
  - Small footprint
  - Robustness
- **Drawback**
  - Quality
- **Major factors for quality degradation [2]**
  - Vocoder
  - Acoustic model → Deep learning
  - Oversmoothing

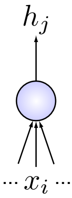
Deep learning [3]

- Machine learning methodology using multiple-layered models
- Motivated by brains, which organize ideas and concepts hierarchically
- Typically artificial neural network (NN) w/ 3 or more levels of non-linear operations



Basic components in NN

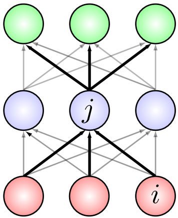
Non-linear unit



$$h_i = f(z_i)$$

$$z_j = \sum_i x_i w_{ij}$$

Network of units



Examples of activation functions

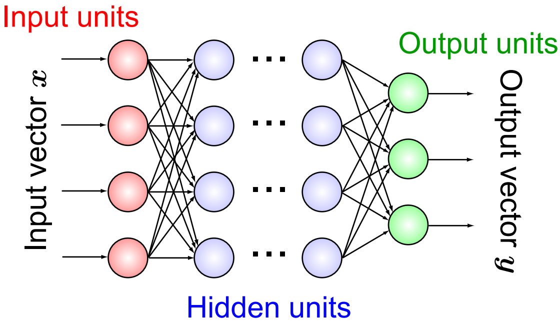
Logistic sigmoid: $f(z_j) = \dfrac{1}{1 + e^{-z_j}}$

Hyperbolic tangent: $f(z_j) = \tanh(z_j)$

Rectified linear: $f(z_j) = \max(z_j, 0)$

Deep architecture

- Logistic regression $\rightarrow$ depth=1
- Kernel machines, decision trees $\rightarrow$ depth=2
- Ensemble learning (e.g., Boosting [4], tree intersection [5]) $\rightarrow$ depth++
- N-layer neural network $\rightarrow$ depth=$N+1$

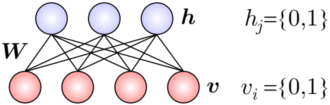


Difficulties to train DNN

- **NN w/ many layers used to give worse performance than NN w/ few layers**
  - Slow to train
  - Vanishing gradients [6]
  - Local minimum

- **Since 2006, training DNN significantly improved**
  - GPU [7]
  - More data
  - Unsupervised pretraining (RBM [8], auto-encoder [9])

Restricted Boltzmann Machine (RBM) [11]



- Undirected graphical model
- No connection between visible & hidden units

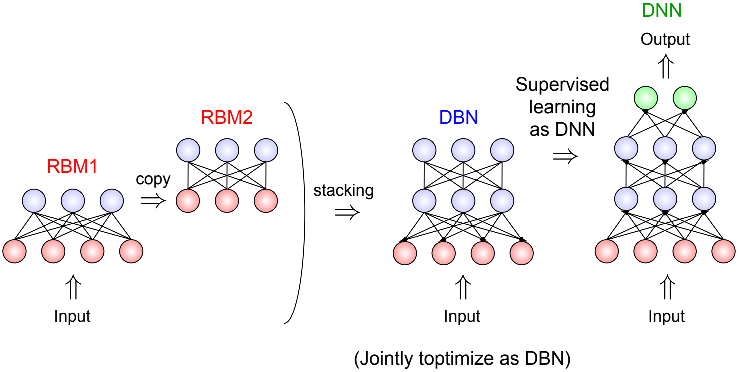
$$p(\boldsymbol{v}, \boldsymbol{h} \mid \boldsymbol{W}) = \frac{1}{Z(\boldsymbol{W})} \exp\left\{-E(\boldsymbol{v}, \boldsymbol{h}; \boldsymbol{W})\right\} \qquad w_{ij}\text{: weight}$$

$$E(\boldsymbol{v}, \boldsymbol{h}; \boldsymbol{W}) = -\sum_i b_i v_i - \sum_j c_j h_j - \sum_{i,j} v_i w_{ij} h_j \qquad b_i, c_j\text{: bias}$$

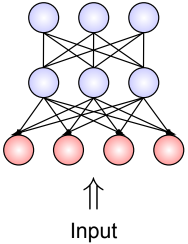
- Parameters can be estimated by contrastive divergence learning [10]

Deep Belief Network (DBN) [8]

- RBMs are stacked to form a DBN
- Layer-wise training of RBM is repeated over multiple layers **(pretraining)**
- Joint optimization as DBN or supervised learning as DNN with additional final layer **(fine tuning)**

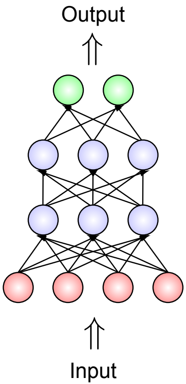


DBN
(feature extractor)



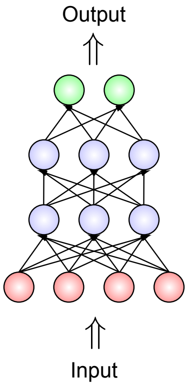
Unsupervised
layer-wise
pre-training

DBN + classification layer
(feature → classifier)



Adding
output layer
(e.g., softmax)

DNN
(feature + classifier)



Supervised
fine-tuning
(backpropagation)

Success of DNN in various machine learning tasks

Tasks

- Vision [12]
- Language
- Speech [13]

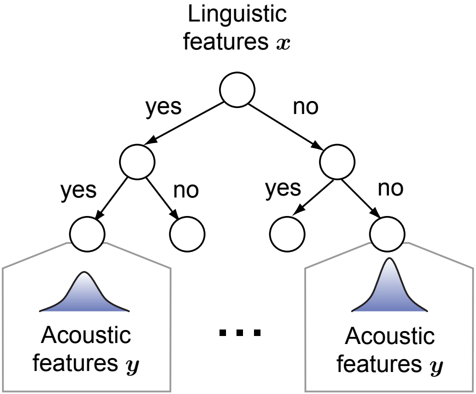
		Word error rates (%)		
Task	Hours of data	HMM-DNN	HMM-GMM w/ same data	HMM-GMM w/ more data
Voice Input	5,870	12.3	N/A	16.0
YouTube	1,400	47.6	52.3	N/A

Products

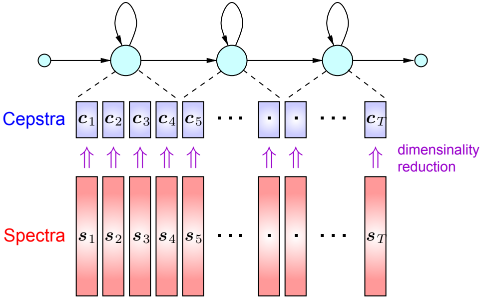
- Personalized photo search [14, 15]
- Voice search [16, 17].

Conventional HMM-GMM [1]

- Decision tree-clustered HMM with GMM state-output distributions



Limitation of HMM-GMM approach (1)
Hard to integrate feature extraction & modeling

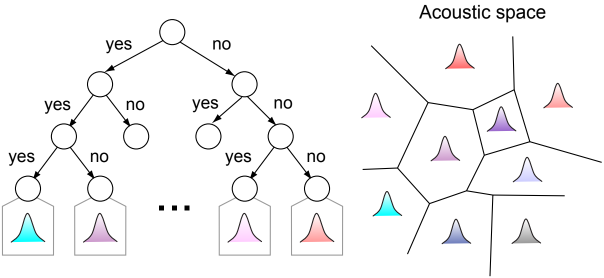


- Typically use lower dimensional approximation of speech spectrum as acoustic feature (e.g., cepstrum, line spectral pairs)
- Hard to model spectrum directly by HMM-GMM due to high dimensionality & strong correlation

→ Waveform-level model [18], mel-cepstral analysis-integrated model [19], STAVOCO [20], MGE-LSD [21]

Limitation of HMM-GMM approach (2)

Data fragmentation



- Linguistic-to-acoustic mapping by decision trees
- Decision tree splits input space into sub-clusters
- Inefficient to represent complex dependencies between linguistic & acoustic features

→ Boosting [4], tree intersection [5], product of experts [22]

Motivation to use deep learning in speech synthesis

- **Integrating feature extraction**
  - Can model high-dimensional, highly correlated features efficiently
  - Layered architecture with non-linear operations offers feature extraction to be integrated with acoustic modeling

- **Distributed representation**
  - Can be exponentially more efficient than fragmented representation
  - Better representation ability with fewer parameters

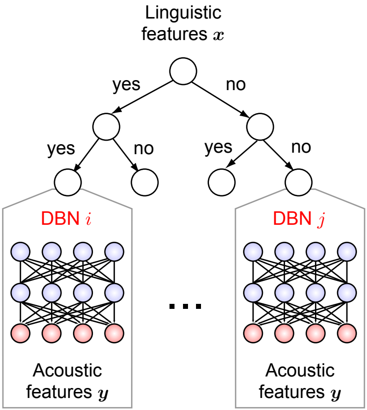
- **Layered hierarchical structure in speech production**
  - concept $\rightarrow$ linguistic $\rightarrow$ articulatory $\rightarrow$ waveform

Deep learning-based approaches

Recent applications of deep learning to speech synthesis

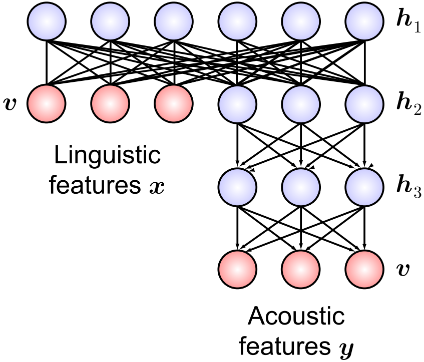
- HMM-DBN (USTC/MSR [23, 24])
- DBN (CUHK [25])
- DNN (Google [26])
- DNN-GP (IBM [27])

HMM-DBN [23, 24]



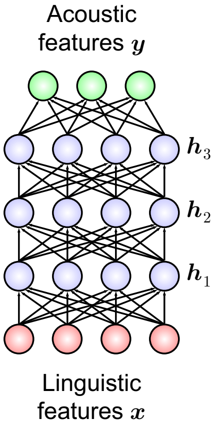
- Decision tree-clustered HMM with DBN state-output distributions
- DBNs replaces GMMs

DBN [25]



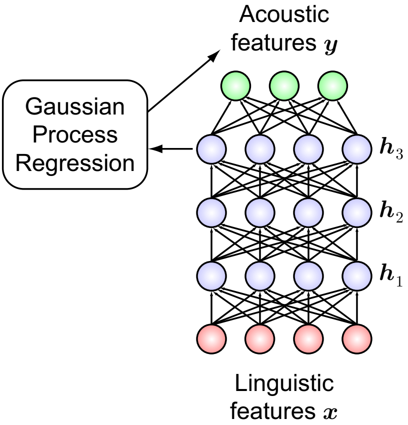
- DBN represents joint distribution of linguistic & acoustic features
- DBN replaces decision trees and GMMs

DNN [26]



- DNN represents conditional distribution of acoustic features given linguistic features
- DNN replaces decision trees and GMMs

DNN-GP [27]



- Uses last hidden layer output as input for Gaussian Process (GP) regression
- Replaces last layer of DNN by GP regression

Comparison

cep: mel-cepstrum, ap: band aperiodicities

$\boldsymbol{x}$: linguistic features, $\boldsymbol{y}$: acoustic features, c: cluster index

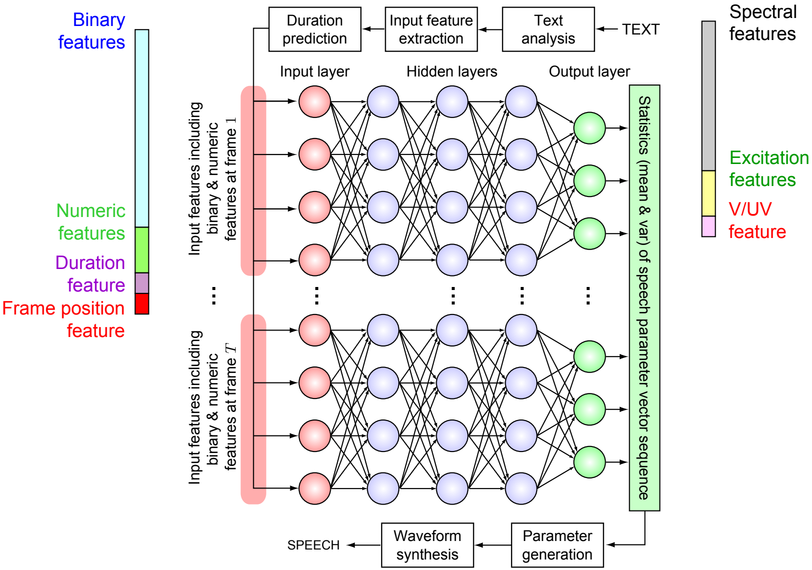
$y \mid x$: conditional distribution of y given x

(y, x): joint distribution between x and y

HMM -GMM	**HMM -DBN**	**DBN**	**DNN**	**DNN -GP**
cep, ap, F_0	spectra	cep, ap, F_0	cep, ap, F_0	F_0
parametric	parametric	parametric	parametric	non-parametric
$\boldsymbol{y} \mid c \leftarrow c \mid \boldsymbol{x}$	$\boldsymbol{y} \mid c \leftarrow c \mid \boldsymbol{x}$	$(\boldsymbol{y}, \boldsymbol{x})$	$\boldsymbol{y} \mid \boldsymbol{x}$	$\boldsymbol{y} \mid \boldsymbol{h} \leftarrow \boldsymbol{h} \mid \boldsymbol{x}$

HMM-GMM is more computationally efficients than others

Framework


Binary features
Numeric features
Duration feature
Frame position feature
Duration prediction
Input feature extraction
Text analysis
TEXT
Input layer
Hidden layers
Output layer
Input features including binary & numeric features at frame 1
Input features including binary & numeric features at frame T
Statistics (mean & var) of speech parameter vector sequence
Spectral features
Excitation features
V/UV feature
SPEECH
Waveform synthesis
Parameter generation

Framework

Is this new? . . . no

- NN [28]
- RNN [29]

What's the difference?

- More layers, data, computational resources
- Better learning algorithm
- Statistical parametric speech synthesis techniques

Experimental setup

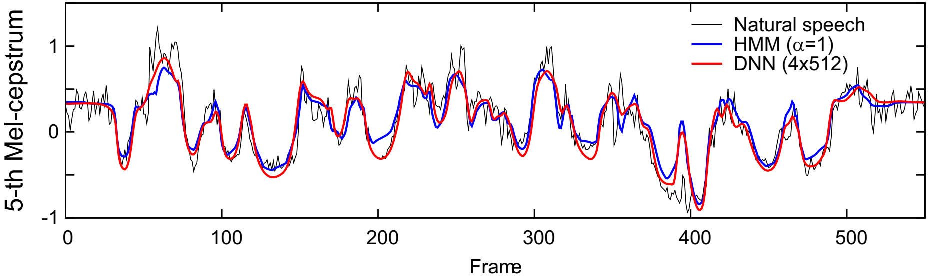
Database	US English female speaker
Training / test data	33000 & 173 sentences
Sampling rate	16 kHz
Analysis window	25-ms width / 5-ms shift
Linguistic features	11 categorical features 25 numeric features
Acoustic features	0–39 mel-cepstrum $\log F_0$, 5-band aperiodicity, Δ, Δ^2
HMM topology	5-state, left-to-right HSMM [30], MSD F_0 [31], MDL [32]
DNN architecture	1–5 layers, 256/512/1024/2048 units/layer sigmoid, continuous F_0 [33]
Postprocessing	Postfiltering in cepstrum domain [34]

Preliminary experiments

- w/ vs w/o grouping questions (e.g., vowel, fricative)
  - Grouping (OR operation) can be represented by NN
  - w/o grouping questions worked more efficiently
- How to encode numeric features for inputs
  - Decision tree clustering uses binary questions
  - Neural network can have numerical values as inputs
  - Feeding numerical values directly worked more efficiently
- Removing silences
  - Decision tree splits silence & speech at the top of the tree
  - Single neural network handles both of them
  - Neural network tries to reduce error for silence
  - Better to remove silence frames as preprocessing

Example of speech parameter trajectories

w/o grouping questions, numeric contexts, silence frames removed

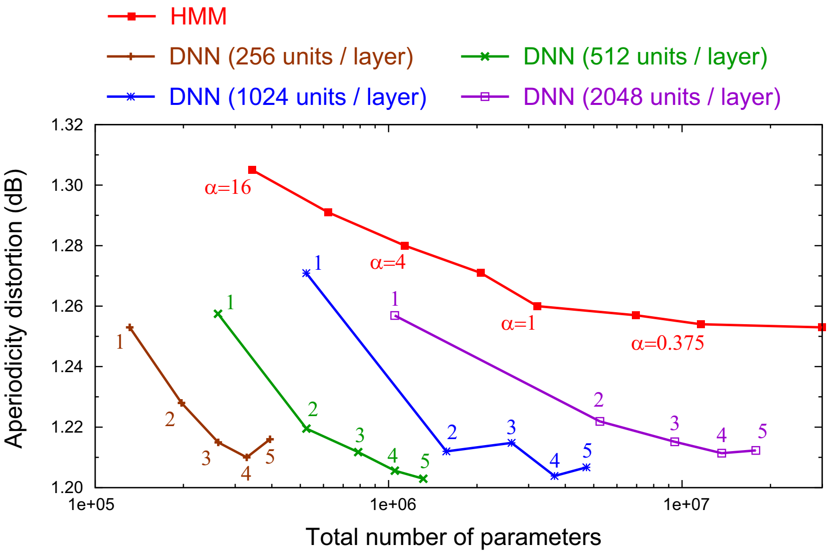


Objective evaluations

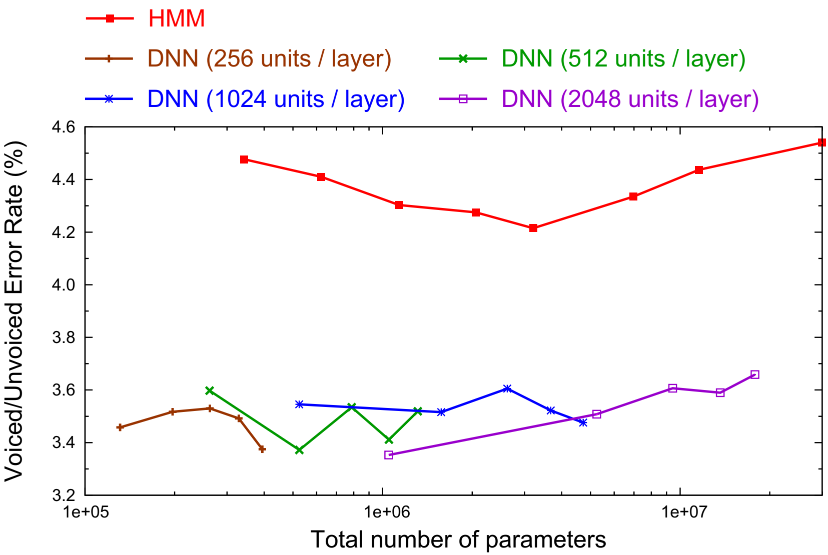
- **Objective measures**
  - Aperiodicity distortion (dB)
  - Voiced/Unvoiced error rates (%)
  - Mel-cepstral distortion (dB)
  - RMSE in $\log F_0$

- **Sizes of decision trees in HMM systems were tuned by scaling (α) the penalty term in the MDL criterion**
  - $\alpha < 1$: larger trees (more parameters)
  - $\alpha = 1$: standard setup
  - $\alpha > 1$: smaller trees (fewer parameters)

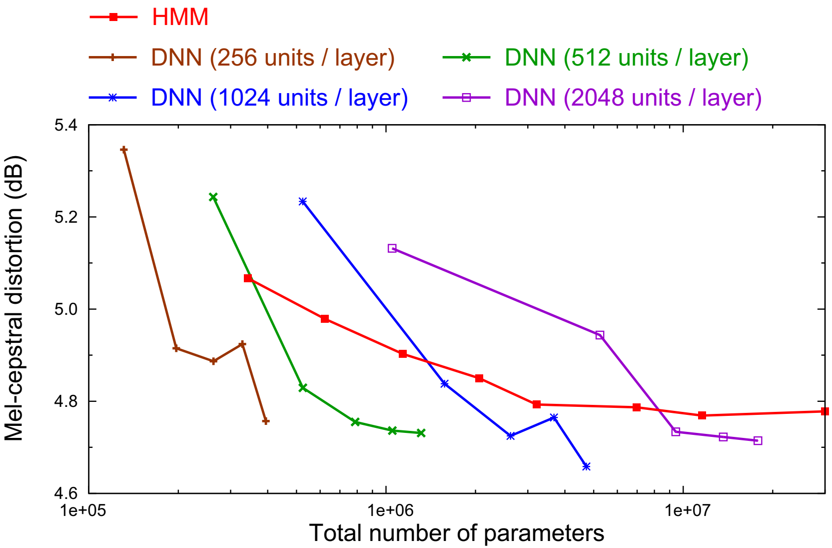
Aperiodicity distortion



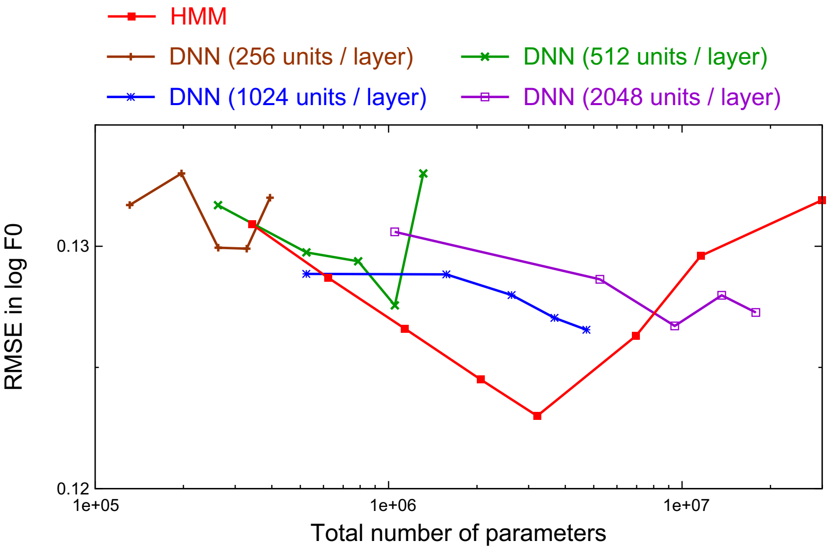
V/UV errors


HMM
DNN (256 units / layer)
DNN (512 units / layer)
DNN (1024 units / layer)
DNN (2048 units / layer)
Voiced/Unvoiced Error Rate (%)
4.6
4.4
4.2
4.0
3.8
3.6
3.4
3.2
1e+05
1e+06
1e+07
Total number of parameters

Mel-cepstral distortion



RMSE in $\log F0$



Subjective evaluations

Compared HMM-based systems with DNN-based ones with similar # of parameters

- Paired comparison test
- 173 test sentences, 5 subjects per pair
- Up to 30 pairs per subject
- Crowd-sourced

HMM (α)	DNN (#layers $\times$ #units)	Neutral	p value	z value
15.8 (16)	**38.5** (4 $\times$ 256)	45.7	$< 10^{-6}$	-9.9
16.1 (4)	**27.2** (4 $\times$ 512)	56.8	$< 10^{-6}$	-5.1
12.7 (1)	**36.6** (4 $\times$ 1 024)	50.7	$< 10^{-6}$	-11.5

Conclusion

Deep learning in speech synthesis

- Aims to replace HMM with acoustic model based on deep architectures
- Different groups presented different architectures at ICASSP 2013
  - HMM-DBN
  - DBN
  - DNN
  - DNN-GP
- DNN-based approach achieved reasonable performance
- Many possible future research topics

References I

[1] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura.
Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis.
In *Proc. Eurospeech*, pages 2347–2350, 1999.

[2] H. Zen, K. Tokuda, and A. Black.
Statistical parametric speech synthesis.
Speech Commun., 51(11):1039–1064, 2009.

[3] Y. Bengio.
Learning deep architectures for AI.
Foundations and Trends in Machine Learning, 2(1):1–127, 2009.

[4] Y. Qian, H. Liang, and F. Soong.
Generating natural F0 trajectory with additive trees.
In *Proc. Interspeech*, pages 2126–2129, 2008.

References II

[5] K. Yu, H. Zen, F. Mairesse, and S. Young.
Context adaptive training with factorized decision trees for HMM-based statistical parametric speech synthesis.
Speech Commun., 53(6):914–923, 2011.

[6] S. Hochreiter, Y. Bengio, P. Frasconi, and J. Schmidhuber.
Gradient flow in recurrent nets: the difficulty of learning long-term dependencies.
In S. Kremer and J. Kolen, editors, *A field guide to dynamical recurrent neural networks*. IEEE Press, 2001.

[7] R. Raina, A. Madhavan, and A. Ng.
Large-scale deep unsupervised learning using graphics processors.
In *Proc. ICML*, volume 9, pages 873–880, 2009.

References III

[8] G. Hinton, S. Osindero, and Y.W. Teh.
A fast learning algorithm for deep belief nets.
Neural Computation, 18(7):1527–1554, 2006.

[9] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol.
Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion.
Journal of Machine Learning Research, 11:3371–3408, 2010.

[10] G.E. Hinton.
Training products of experts by minimizing contrastive divergence.
Neural Computation, 14(8):1771–1800, 2002.

References IV

[11] P Smolensky.

Information processing in dynamical systems: Foundations of harmony theory.

In D. Rumelhard and J. McClelland, editors, *Parallel Distributed Processing*, volume 1, chapter 6, pages 194–281. MIT Press, 1986.

[12] A. Krizhevsky, I. Sutskever, and G. Hinton.

Imagenet classification with deep convolutional neural networks.

In *Proc. NIPS*, pages 1106–1114, 2012.

[13] G. Hinton, L. Deng, D. Yu, G. Dahl, A.-R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury.

Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups.

IEEE Signal Processing Magazine, 29(6):82–97, 2012.

References V

[14] C. Rosenberg.
Improving photo search: a step across the semantic gap.
http://googleresearch.blogspot.co.uk/2013/06/improving-photo-search-step-across.html.

[15] K. Yu.
https://plus.sandbox.google.com/103688557111379853702/posts/fdw7EQX87Eq.

[16] V. Vanhoucke.
Speech recognition and deep learning.
http://googleresearch.blogspot.co.uk/2012/08/speech-recognition-and-deep-learning.html.

References VI

[17] Bing makes voice recognition on Windows Phone more accurate and twice as fast.

http://www.bing.com/blogs/site_blogs/b/search/archive/2013/06/17/dnn.aspx.

[18] R. Maia, H. Zen, and M. Gales.

Statistical parametric speech synthesis with joint estimation of acoustic and excitation model parameters.

In *Proc. ISCA SSW7*, pages 88–93, 2010.

[19] K. Nakamura, K. Hashimoto, Y. Nankaku, and K. Tokuda.

Integration of acoustic modeling and mel-cepstral analysis for HMM-based speech synthesis.

In *Proc. ICASSP*, pages 7883–7887, 2013.

References VII

[20] T. Toda and K. Tokuda.
Statistical approach to vocal tract transfer function estimation based on factor analyzed trajectory hmm.
In *Proc. ICASSP*, pages 3925–3928, 2008.

[21] Y.-J. Wu and K. Tokuda.
Minimum generation error training with direct log spectral distortion on LSPs for HMM-based speech synthesis.
In *Proc. Interspeech*, pages 577–580, 2008.

[22] H. Zen, M. Gales, Y. Nankaku, and K. Tokuda.
Product of experts for statistical parametric speech synthesis.
IEEE Trans. Audio Speech Lang. Process., 20(3):794–805, 2012.

References VIII

[23] Z.-H. Ling, L. Deng, and D. Yu.
Modeling spectral envelopes using restricted Boltzmann machines for statistical parametric speech synthesis.
In *Proc. ICASSP*, pages 7825–7829, 2013.

[24] Z.-H. Ling, L. Deng, and D. Yu.
Modeling spectral envelopes using restricted Boltzmann machines and deep belief networks for statistical parametric speech synthesis.
IEEE Trans. Audio Speech Lang. Process., 21(10):2129–2139, 2013.

[25] S. Kang, X. Qian, and H. Meng.
Multi-distribution deep belief network for speech synthesis.
In *Proc. ICASSP*, pages 8012–8016, 2013.

References IX

[26] H. Zen, A. Senior, and M. Schuster.
Statistical parametric speech synthesis using deep neural networks.
In *Proc. ICASSP*, pages 7962–7966, 2013.

[27] R. Fernandez, A. Rendel, B. Ramabhadran, and R. Hoory.
F0 contour prediction with a deep belief network-Gaussian process hybrid model.
In *Proc. ICASSP*, pages 6885–6889, 2013.

[28] O. Karaali, G. Corrigan, and I. Gerson.
Speech synthesis with neural networks.
In *Proc. World Congress on Neural Networks*, pages 45–50, 1996.

[29] C. Tuerk and T. Robinson.
Speech synthesis using artificial network trained on cepstral coefficients.
In *Proc. Eurospeech*, pages 1713–1716, 1993.

References X

[30] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura.
A hidden semi-Markov model-based speech synthesis system.
IEICE Trans. Inf. Syst., E90-D(5):825–834, 2007.

[31] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi.
Multi-space probability distribution HMM.
IEICE Trans. Inf. Syst., E85-D(3):455–464, 2002.

[32] K. Shinoda and T. Watanabe.
Acoustic modeling based on the MDL criterion for speech recognition.
In *Proc. Eurospeech*, pages 99–102, 1997.

[33] K. Yu and S. Young.
Continuous F0 modelling for HMM based statistical parametric speech synthesis.
IEEE Trans. Audio Speech Lang. Process., 19(5):1071–1079, 2011.

References XI

[34] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura.
Incorporation of mixed excitation model and postfilter into HMM-based text-to-speech synthesis.
IEICE Trans. Inf. Syst., J87-D-II(8):1563–1571, 2004.