Has DNN Outperformed HMM in Speech Synthesis?

Zhehuai Chen

Speech Lab
Department of Computer Science and Engineering
Shanghai Jiao Tong University

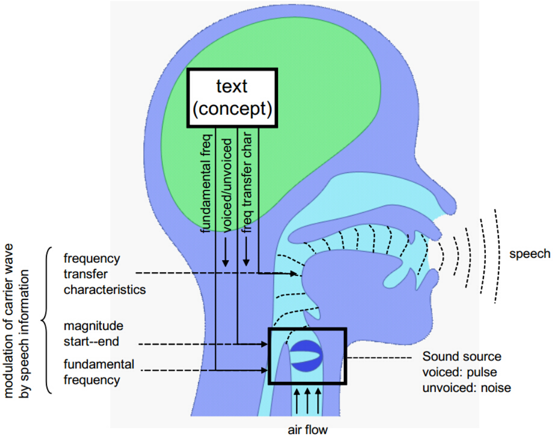
Oct. 2014

Outline

- Introduction
- DNN-based Speech Synthesis System Implementation
- System Performance Analysis
- Experiment Results
- Conclusions

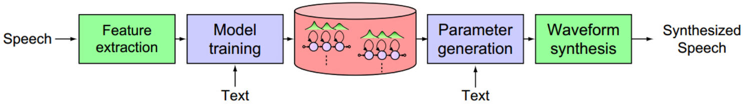
Introduction

Typical Speech Synthesis Flow


text (concept)
fundamental freq
voiced/unvoiced
freq transfer char
modulation of carrier wave by speech information
frequency transfer characteristics
magnitude start--end
fundamental frequency
speech
Sound source voiced: pulse unvoiced: noise
air flow

Introduction

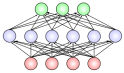
HMM-based Speech Synthesis



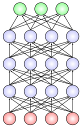
- To map from input linguistic feature to output acoustic features for synthesis
- Hidden Markov model (HMM) as its acoustic model

Introduction

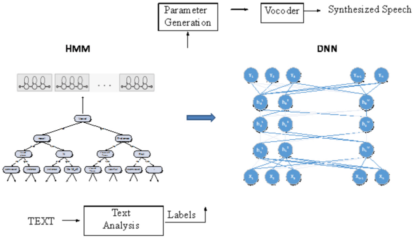
Deep Neural Network in Speech Synthesis



Shallow Neural Network



Deep Neural Network (DNN)



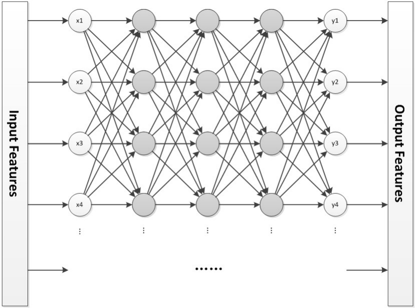
- suitable to model a long-span, intricate transform compactly with a deep-layered structure
- successfully used in speech recognition, also applied to speech synthesis (Zen, et al) to replace the HMM in the system

DNN-based Speech Synthesis System Implementation

Framework of DNN-based System

- 3-hidden-layer Nearual Network between linguistic full context labels and acoustic waveform parameters
- Rich contexts packed into a long vector frame-by-frame are used as input feature
- The input features are mapped to output acoustic features by a trained DNN using forward propagation, dynamic features included

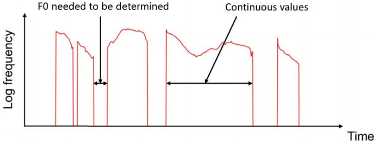
- Vocoding Process is the same to HMM-based system



System Performance Analysis

Different Aspects between DNN & HMM system

- Model Framework
  Long-span and highly-complex $vs.$ shallow but carefully-designed
- Data Useage
  Training Model using all data $vs.$ Data Clustering
- F0 Modeling
  Continuous F0 modelling $vs.$ traditional Multi-space Probability Distribution(MSD-HMM)



Experiments

DNN System Setup

- Training Data
  A U.S. female English speaker, slt and a U.S. male English speaker, awb. Split into Training Set & Test Set.
- Framework of System
  3 hidden layers and each with 1024 nodes. mini-batch=256. Modified version of TNet as the training tool.
- Acoustic Feature Setup
  Continuous F0 modeling using Interpolation, 24 Mel-Cepstral spectral coefficients, 5 Band Aperiodic Components

Experiments

System Comparison

- MSD-HMM vs. CF-HMM vs. DNN

System model	Female			Male		
	RMSE	VCE (%)	MSD	RMSE	VCE (%)	MSD
MSD-HMM	16.02	5.24	0.20	15.11	3.52	0.18
CF-HMM	10.56	6.51	0.20	12.17	4.77	0.18
DNN	12.40	6.27	0.22	13.26	4.96	0.17

Table: objective measures of different speech synthesis system

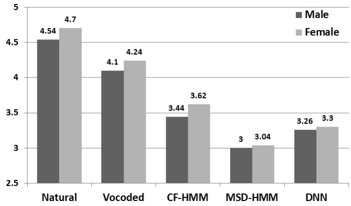
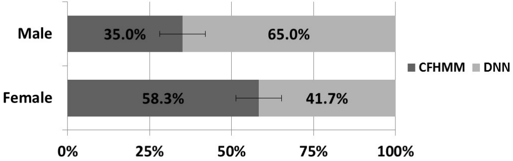
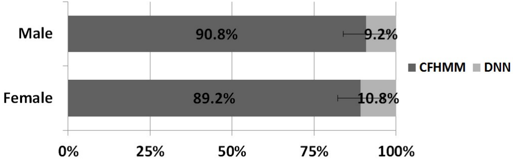


Figure: subjective measures of different speech synthesis system

- F0 modeling ability analysis



- Spectrum modeling ability analysis



Conclusion

- CF-HMM system should be taken as the baseline to compare with DNN-based system, Because of its more similar input features and output features with DNN-based system.
- The ability of F0 modelling is similar between 2 systems, while CF-HMM system performs better in spectrum.
- No enough evidence shows that the modeling ability of a hierarchical complex structure has outperformed that of a shallow but carefully-designed and optimized one.

So how we can analyze the modeling ability and proficiency between them and then realize these potentials is a topic for future investigation.