DNN-Based Duration Modeling for Synthesizing Short Sentences

Péter Nagy, Géza Németh

{nagyp, nemeth}@tmit.bme.hu

Department of Telecommunications and Media Informatics

Budapest University of Technology and Economics

MŰEGYETEM 1782

SmartLab
Intelligent Interactions

Introduction

- High quality, intelligible artifical speech
- Naturalness of synthetic speech below the levels of human speech
  - Problems with the generated synthetic prosody
  - Key aspect: duration
- Statistical parametric speech synthesis
  - Hidden Markov model (HMM) based approach
    - Context dependent decision tree clustered hidden semi Markov models with Gaussian distributions
    - Multi-level duration models
  - Deep netural network (DNN) based approach
    - Feed-forward neural network for duration prediction

Short sentences

- Sentences with one, two or three syllables
- Main focus of this study
- Phone durations are context dependent
  - Dependent on word and utterance length
  - Proper phone durations improve intelligibility and naturalness

- HMMs underperform in these cases
  - Intelligibility highly degraded due to the state-level inherent averaging

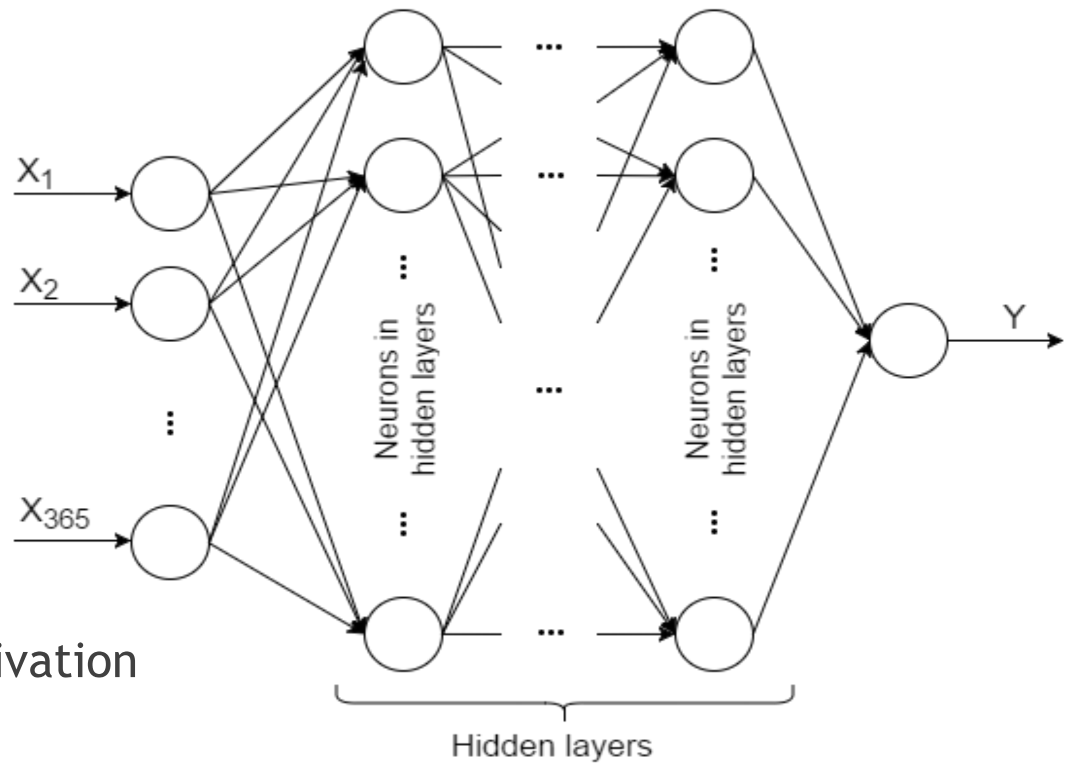
Database specifications

- The Hungarian Parallel Precision Speech Database (PPSD) corpus
- Recordings from 14 speakers (7 female, 7 male)
- 1992 phonetically balanced sentences from different novels per speaker
- Additional 522 utterances
  - Contains interrogative and short sentences
- ~3 hours of speech per speaker on average
- Corpus covers all possible different phoneme transitions
- Annotated and segmented by automatic methods and refined manually
- 2 speakers were selected (1 female, 1 male)

HMM Training

- Hungarian derivative of HTS 2.3beta
- Baseline system
- 3 voices per speaker
  - HMM-NO: Speaker adapted, 500 normal length utterances
  - HMM-SH: Speaker adapted, 400 short and 100 normal length utterances
  - HMM-SI: Speaker dependent voice, 2300 utterances
- Features
  - 39 mel-cepstral coefficients (including the 0th coefficient)
  - $\log(F_0)$
  - Aperiodicity measures with dynamic features

DNN Training



- Adadelta optimization
- In hidden layers PReLUs as activation
- Output layer: linear activation
- Orthogonal weight initialization between hidden layers
- Glorot weight initialization between input-hidden and hidden-output layers
- To avoid feature co-adaptation dropout with 50% probability
- Early stopping set to 50 epochs

DNN Training Parameters

Feature type	Feature	#	Type
Input	Quinphone	5*68	One-hot
	Forward/backward position of actual phoneme/syllable/word/phrase in syllable/word/phrase/sentence	4*2	Numeric
	Number of phonemes/syllables/words/phrases in the previous/current/next syllable/word/phrase/sentence	4*3	Numeric
	Number of phonemes/syllables/words in the current sentence	3	Numeric
	The previous/current/next phoneme is a vowel of a short sentence	3	Binary
Total number of input features: 366			
Output	Duration	1	Continuous
Total number of output features: 1			

Nagy, P., Németh, G., DNN-Based Duration Modeling for Synthesizing Short Sentences

Evaluation

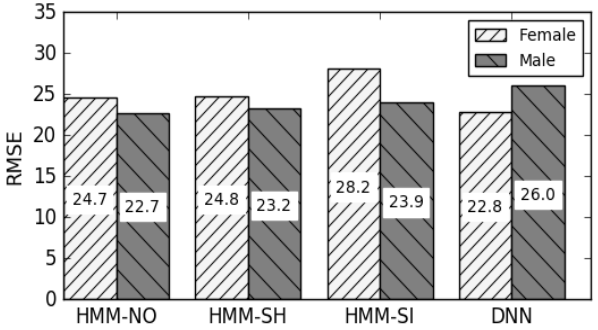
- Hyperparameter optimization with manual grid search
- Optimized parameters: number of hidden layers, number of neurons, minibatch size
- 89 training cycles with the female voice, 74 cycles with the male voice

Voice	# of Layers	# of Neurons	Minibatch	Epochs	MSE
Female	7	900	128	292	0.0029671
	5	1024	128	230	0.0030813
	5	1800	64	317	0.0030924
	7	2048	128	142	0.0031296
Male	7	750	64	126	0.0030007
	5	2048	128	147	0.0030062
	3	1024	64	65	0.0030277
	5	1024	128	230	0.0030813

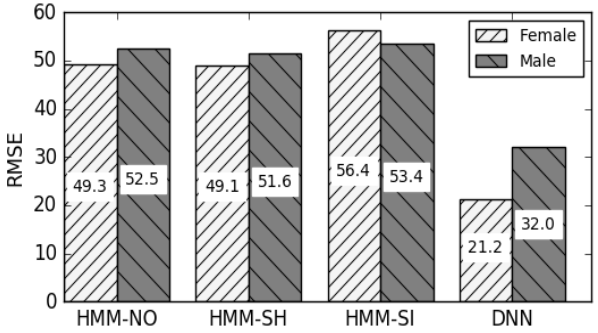
RMSE and Correlation

- RMSE

Normal

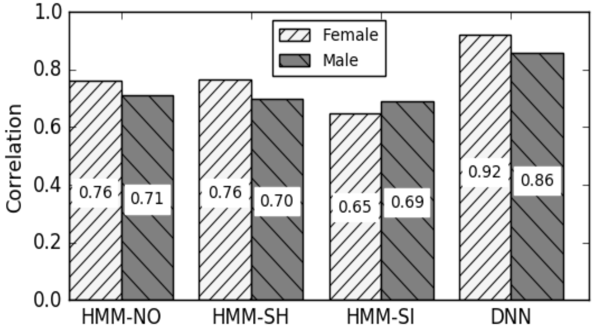


Short

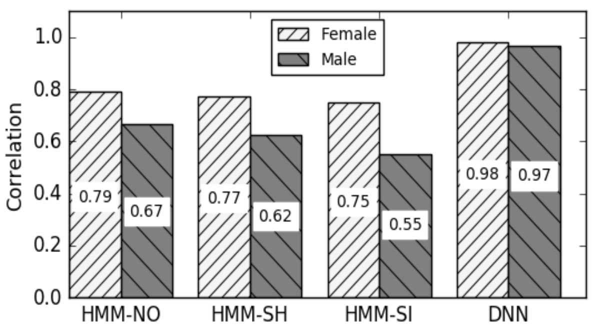


- Correlation

Normal



Short



Nagy, P., Németh, G., DNN-Based Duration Modeling for Synthesizing Short Sentences

Mean durations

- Inverse proportion between syllable count and phoneme durations

			Natural speech	HMM-NO	HMM-SH	HMM-SI	DNN
Female	Normal	V	101.9	103.3	104.5	101.1	109.6
		C	70.6	70	71.1	69.8	74.7
	Short	V	176.7	138.3	148.5	137.7	174.9
		C	114.9	95.5	97.5	85.1	113.9
Male	Normal	V	85.1	82.6	83.7	82.2	96.1
		C	65.5	65.3	65.8	67.9	74.6
	Short	V	153	105.2	111.3	122.8	162.4
		C	101.6	83.5	86.4	93.8	109.9

Summary

- DNN-based duration prediction using FFNN
- The selected contextual features are suitable for prediction
- DNNs can reach the modeling performance of HMMs
  - And can outperform HMMs in case of short sentences
  - Lower prediction error, higher correlation
- Future plans
  - Conduct subjective listening tests
  - Sequential nature of speech is ignored
    - LSTM architecture
  - Introduce additional contextual features