Ensemble Deep Neural Network Based Speech Synthesis

Bálint Pál Tóth, Kornél István Kis, György Szaszák, Géza Németh

Department of Telecommunications and Media Informatics
Budapest University of Technology and Economics

SPECOM 2016

24.08.2016.

SmartLab
Intelligent Interactions

http://smartlab.tmit.bme.hu

NVIDIA GPU EDUCATION CENTER

Introduction

Nowadays statistical paramteric TTS systems mostly use the transcriptions (language rules) only to estimate prosodic stress

Like open-source solutions

In most cases, no extra care for prosodic stress is present in the TTS system

In this paper, F0 is investigated only

Problem statement

1. F0 trajectories generated by a single Deep Neural Network show reduced performance regarding prosodic stress
   - The different stress levels are averaged
   - A single input (stress level) is not sufficient to have enough effect.
   - An ensemble might work better for sparse modeling

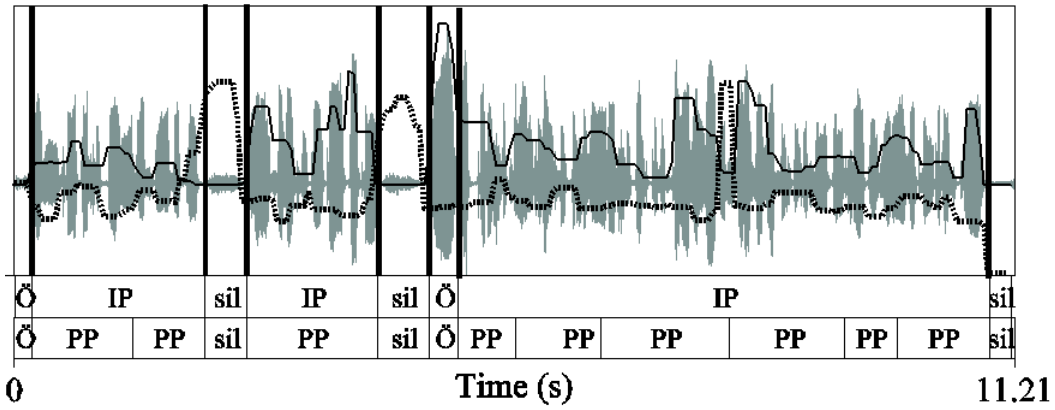
2. Syntactic prosodic stress has poor correlation with the actual stress model of the speaker.

Main Concept 1
Introduction of ensembles

Ensemble of Deep Neural Networks instead of a single deep neural net.

→Expected to result in more focus on prosodic stress.

The stress annotation in the *training* corpora is obtained using the waveform itself.



Main Concept 2
Stress annotation

In the proposed solution – 4 stress levels extracted from the waveform

```
{0} : no stress
{1} : low intensity stress
{2} : high intensity stress
{X} : unknown intensity
```

Unknown intensity usually refers to silent periods at the beginning and at the end of the utterances

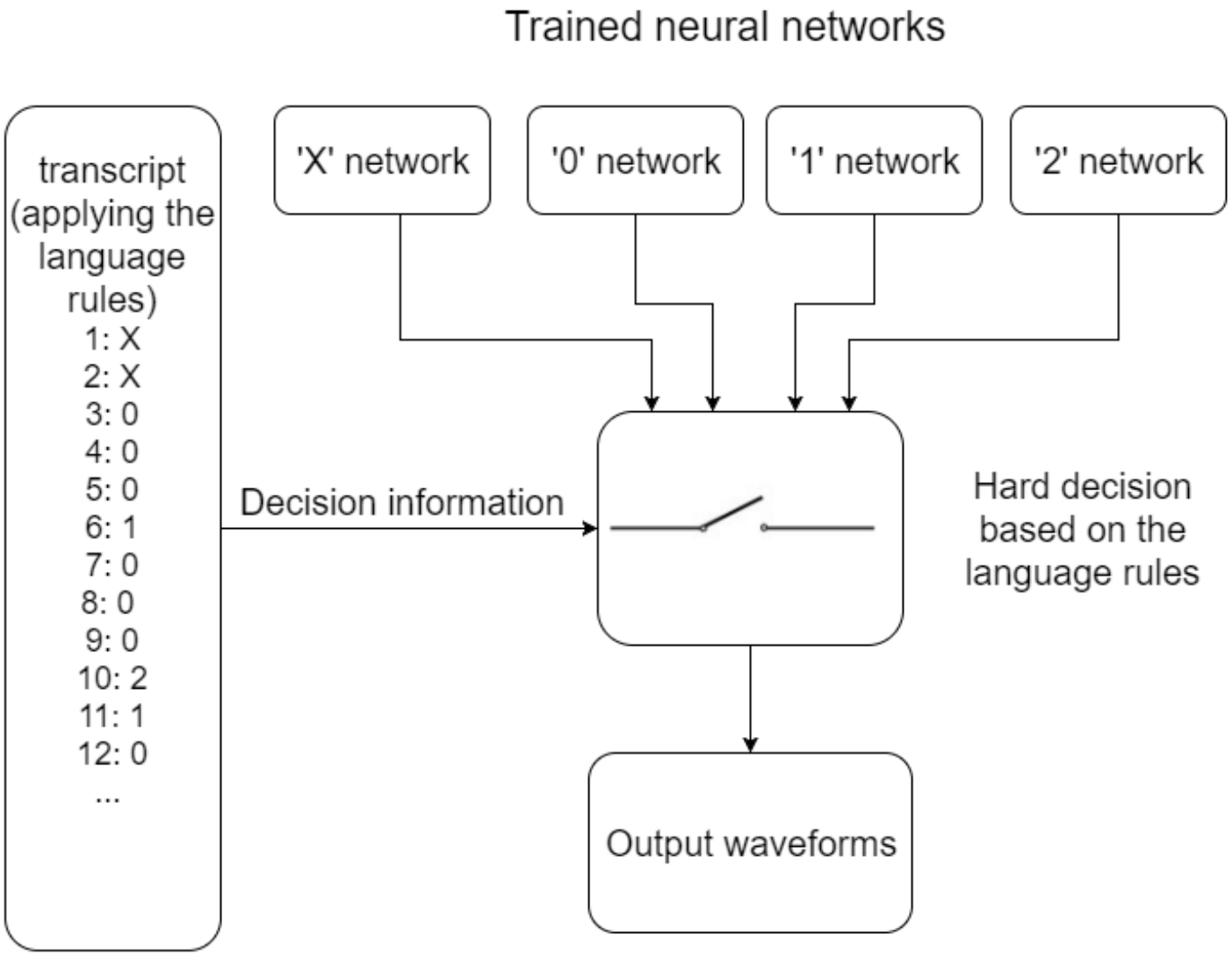
In the synthesis phase stress is annotated by language rules.

Main Concept 3
Training and Evaluation

Four feedforward deep neural networks (DNNs) were trained, each refers to one stress level.

For evaluation purpose the we compare the ensemble with the single DNN model.

Generating output waveforms



Training database

Cca. 2000 declarative Hungarian sentences (from *Hungarian precisely labeled and segmented, parallel speech database*).

Quinphone model + 25 numerical features (*eg. number of syllables in the current word*).

Input is scaled to have zero mean and unit variance, output is minmax scaled between 0.01-0.99.

Training details

Best single model contains *3 hidden layers with 3000 neurons* per layer.

- Input: 363
- Hidden layers: 3000-3000-3000
- Output: 2 (F0, V/UV)

ReLU was used in each hidden layers, the output is sigmoid-equipped.

30% *dropout* and *early stopping (50)* was used with *mini-batch gradient descent with Nesterov momentum.*

Results

Objective and subjective evaluation were carried out

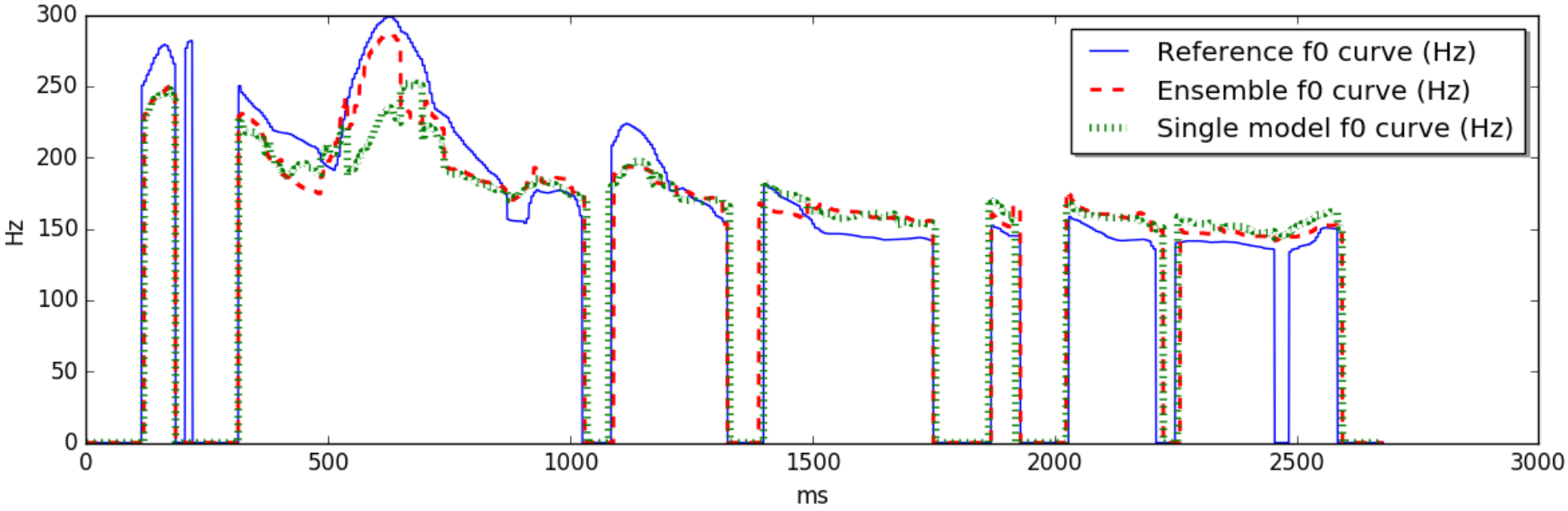
Objective evaluation:

Pearson correlation was calculated between

- The reference and the single model output
- The reference and the ensemble model output

61% of the samples had a slightly worse Pearson value in the ensemble model

Results
An example F0 trajectory

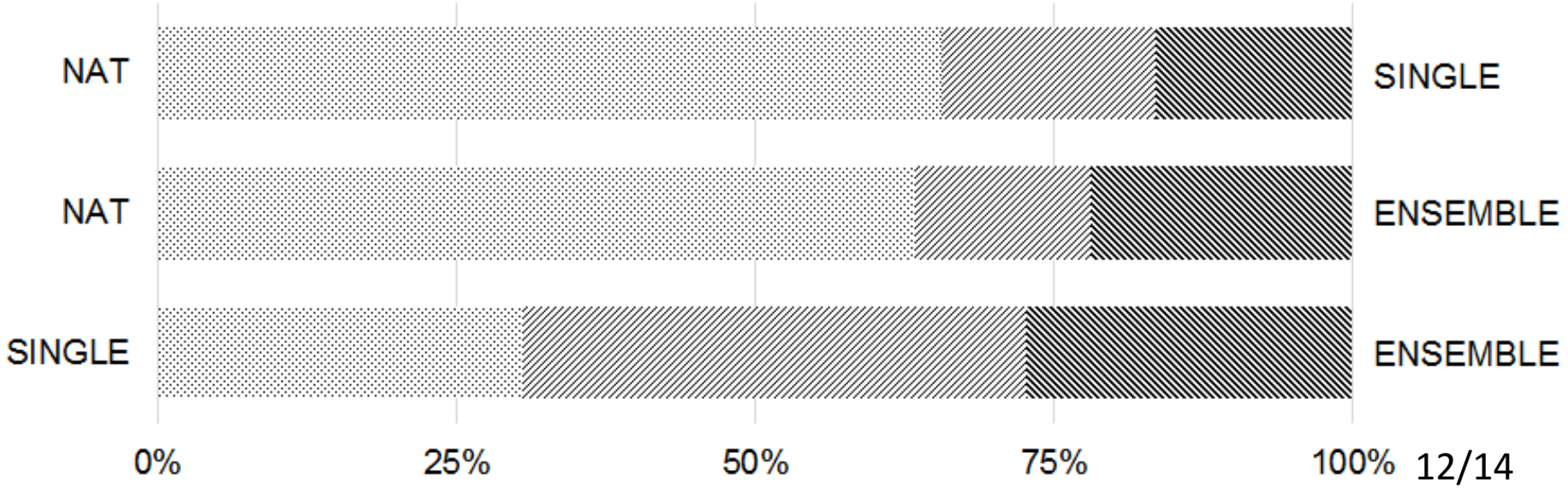


Results
CMOS listening test

72 utterances / test subject, 16 listeners, 22-70 years old

Systems:

- Natural utterances (**NAT**),
- vanilla stress model with single DNN (**SINGLE**),
- proposed stress model with ensemble of DNNs (**ENSEMBLE**).

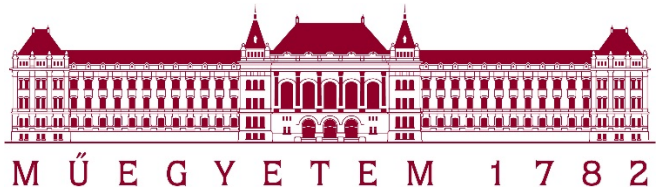


Conclusions and discussion

An appropriate ensemble of networks can outperform a single model.

Ensemble approach is useful when dealing with sparse modeling.

A waveform-driven stress approach causes a noticeable improvement in performance.



BME TMIT

Thank you for your attention!

Questions are welcome

SmartLab
Intelligent Interactions

http://smartlab.tmit.bme.hu

NVIDIA
GPU
EDUCATION
CENTER