# Rozpoznávání fonémů pomocí neuronových sítí

Martin Majer

Katedra kybernetiky
Fakulta aplikovaných věd
Západočeská univerzita v Plzni

# Úvod do problematiky a cíle práce

- porovnání různých typů neuronových sítí a parametrizací řečového signálu pro úlohu rozpoznávání fonémů
- porovnáváno na dvou datových sadách v českém jazyce:
  - ŠkodaAuto - 47 řečníků, 14523 nahrávek
  - SpeechDat-E - 924 řečníků, 39560 nahrávek
- foném = nejmenší lingvistická jednotka schopná rozlišovat významové jednotky (slova)
- rozpoznávání fonémů = úloha, jejíž cílem je pro danou zvukovou nahrávku získat odpovídající sekvenci fonémů

# Typy příznaků

- využity příznaky ve frekvenční oblasti:
  - logaritmované energie banky filtrů
  - mel-frekvenční kepstrální koeficienty
- využití Z-score normalizace

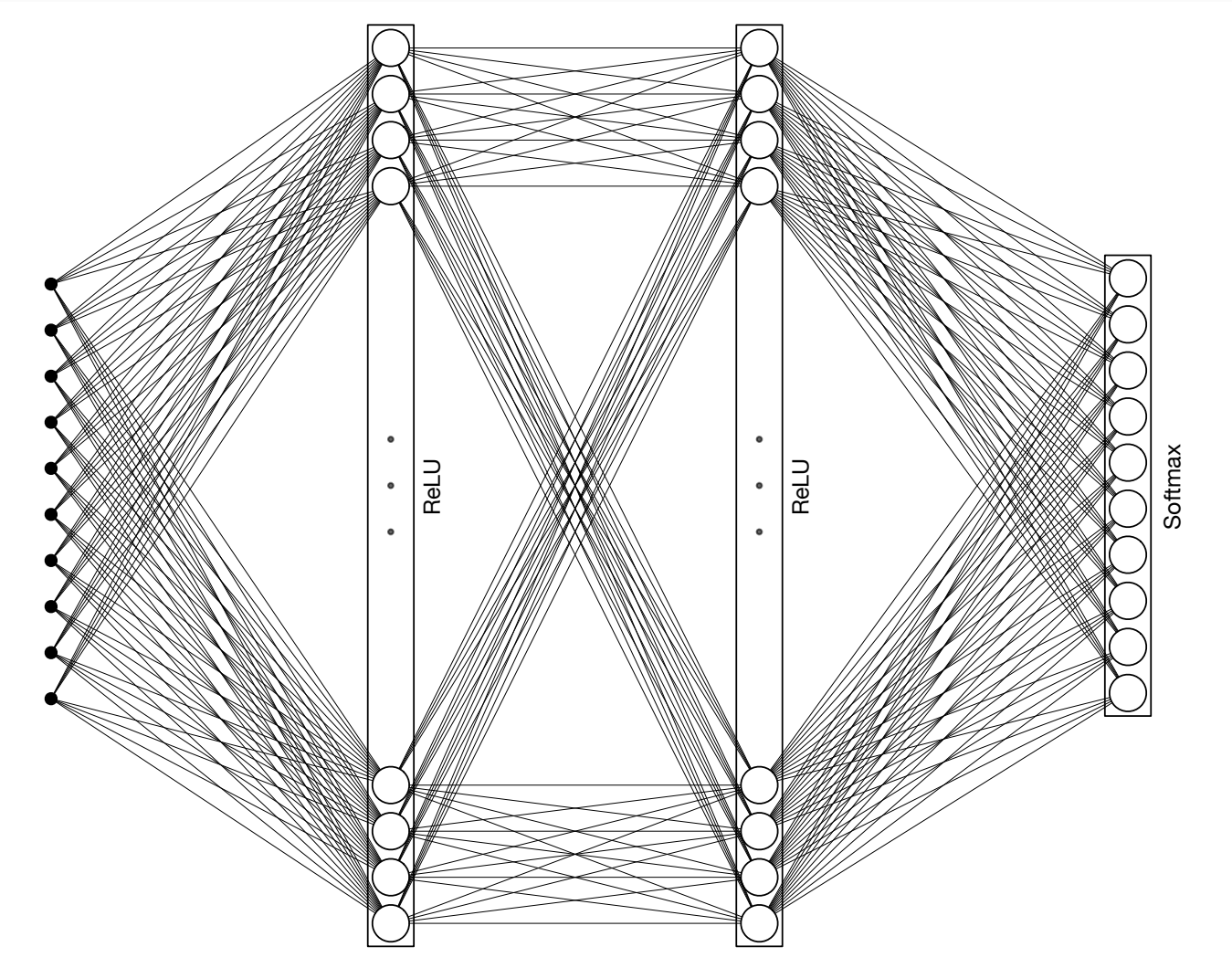
$$z = \frac{x - \mu}{\sigma}$$

- využití delta a delta-delta koeficientů

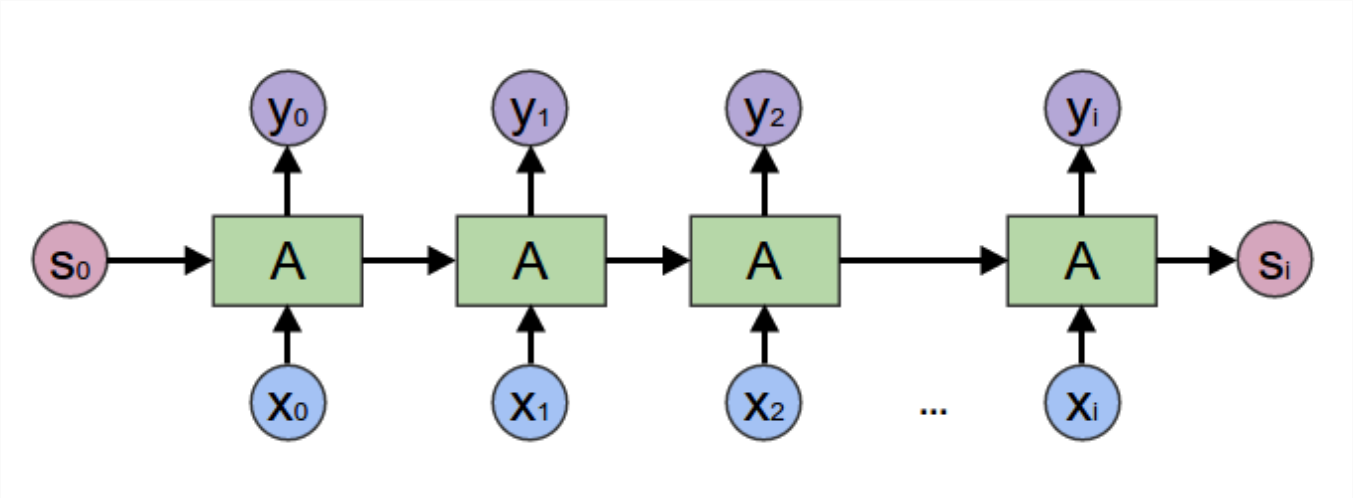
# Typy neuronových sítí

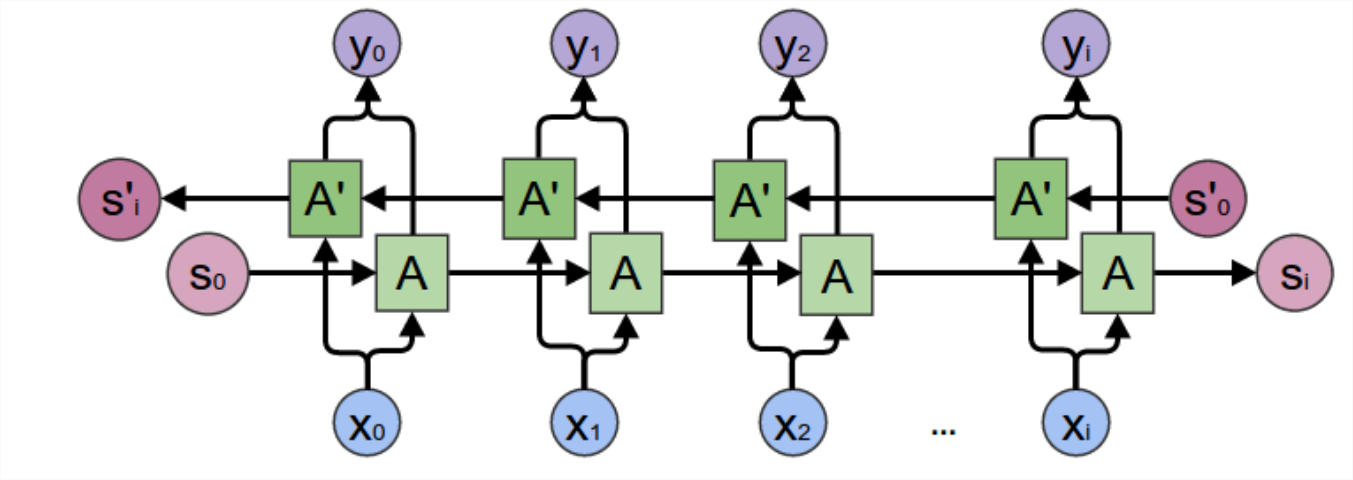
- využité typy neuronových sítí:
  - dopředná neuronová síť s Viterbiho dekodérem
  - LSTM/GRU/obousměrná LSTM s Viterbiho dekodérem
  - LSTM/obousměrná LSTM s CTC
- Viterbiho dekodér využívá zerogramový jazykový model
- využita metoda předčasného ukončení a zašumění trénovacích dat Gaussovským šumem

# Dopředná neuronová síť - schéma



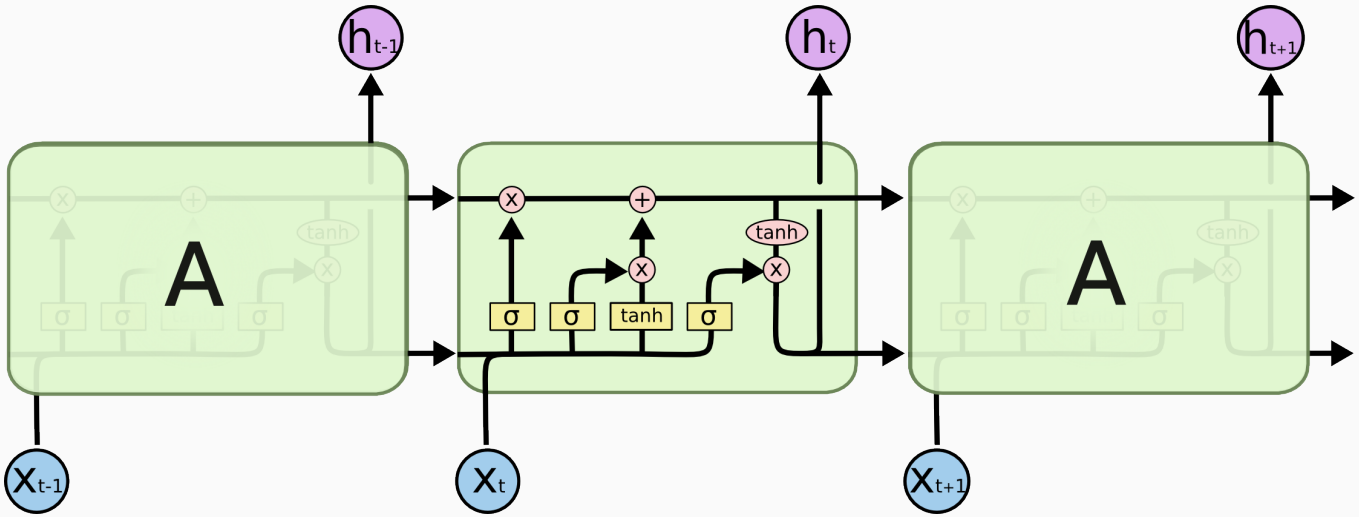
# Rekurentní neuronová síť - schéma




Převzato z *http://colah.github.io/posts/2015-09-NN-Types-FP*

# LSTM - schéma

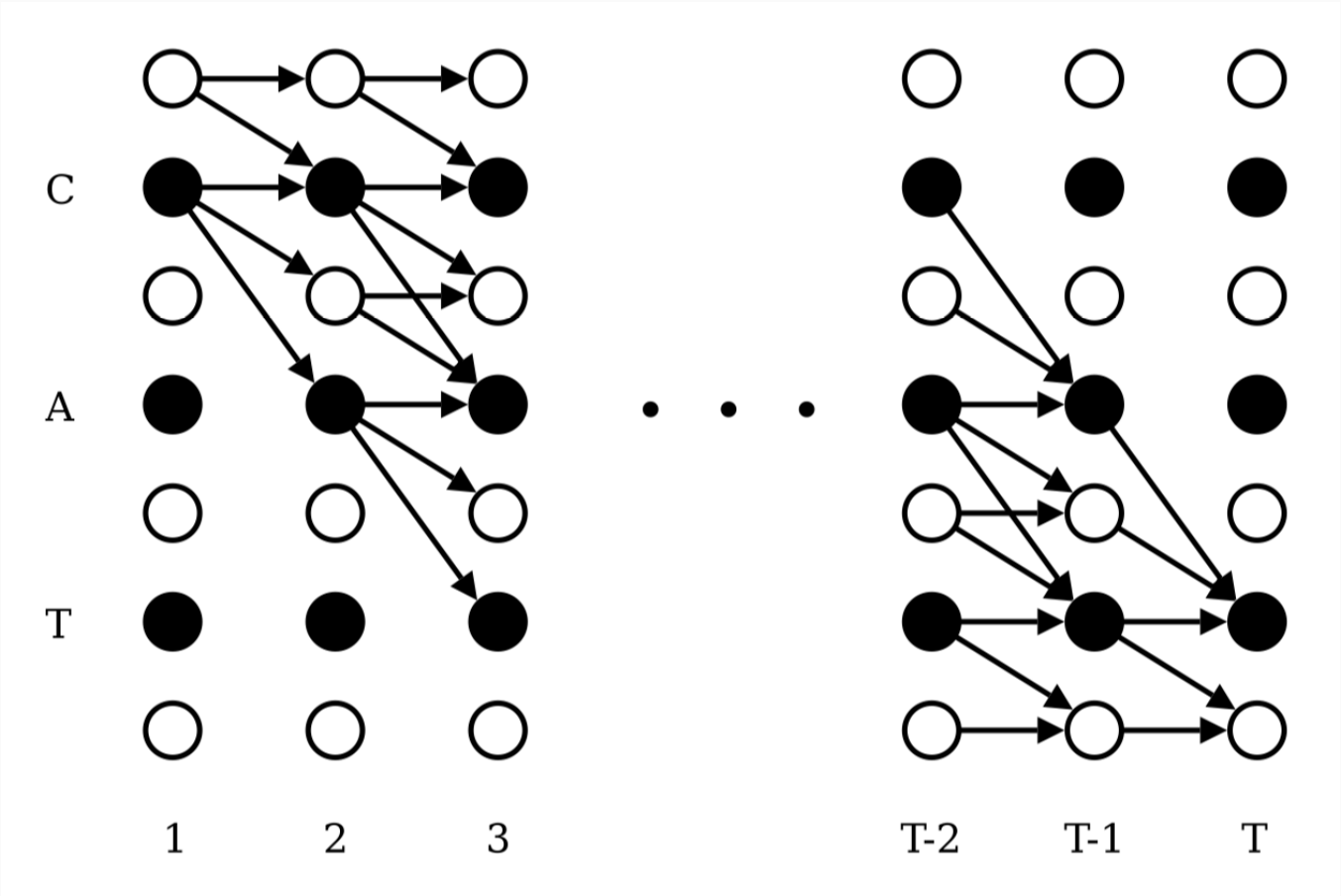



Převzato z *http://colah.github.io/posts/2015-08-Understanding-LSTMs*

# CTC (connectionist temporal classification)

- není potřeba segmentovat nahrávky po fonémech
- není potřeba dále zpracovávat výstup sítě
- hledáme nejpravděpodobnější značkování $l$ pro vstupní obraz $x$, tj. $\mathit{argmax}\ p(l \mid x) \rightarrow$ zavedení prázdného znaku a využití dopředného a zpětného algoritmu
- využito dekódování nejlepší cesty bez jazykového modelu

# CTC - schéma




Převzato z *Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks, A. Graves, 2006*

# Vyhodnocení

- hlavní vyhodnocovací metrika:
  - **phoneAcc [%]** - přesnost modelu po dekódování

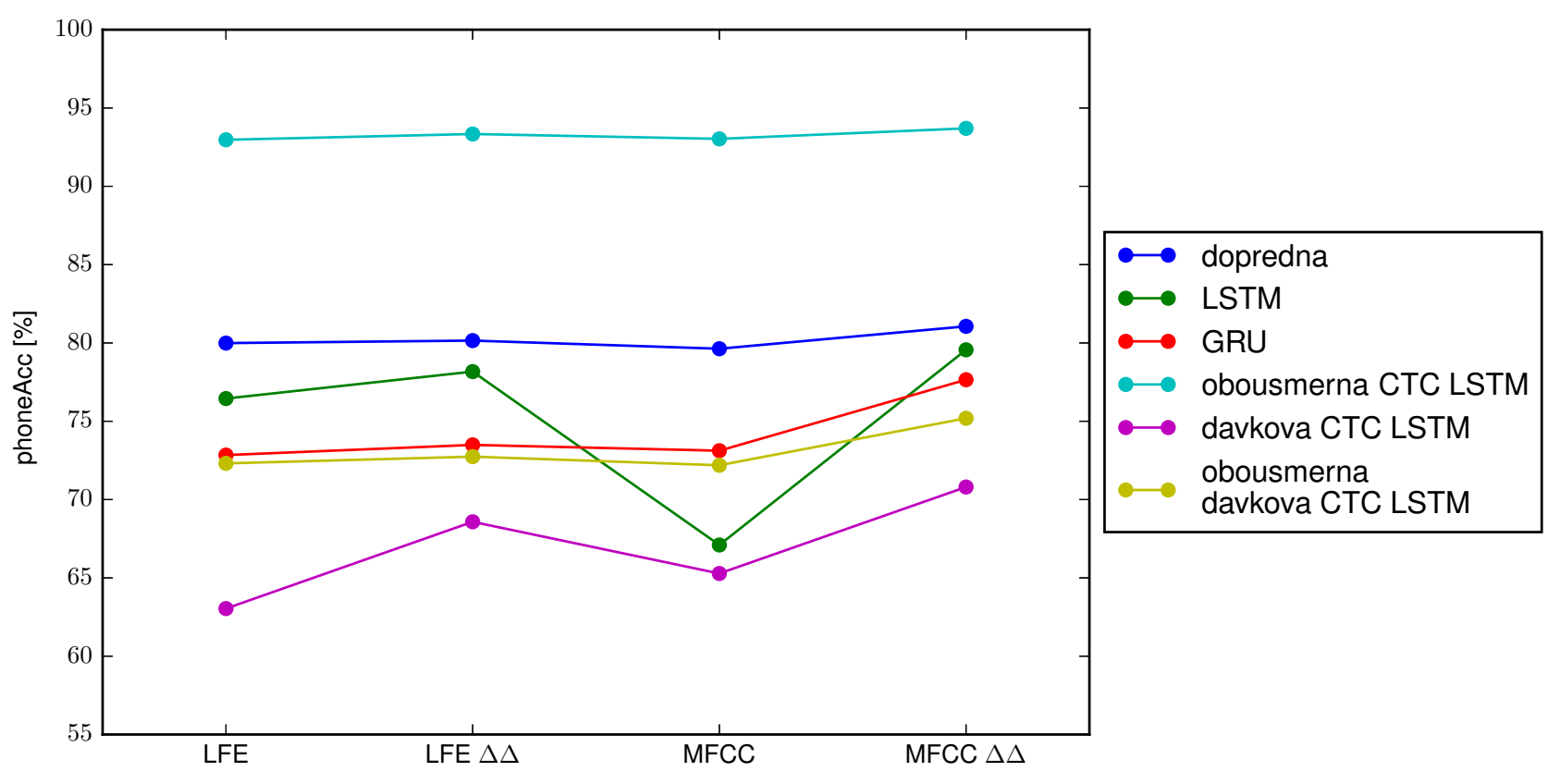
$$\text{phoneAcc} = \frac{\textit{phones correct} - \textit{phones inserted}}{\textit{phones total}} \cdot 100$$

- další zohledněné metriky:
  - **frameAcc [%]** - přesnost klasifikace modelu vyhodnocována po jednotlivých segmentech nahrávek
  - **phoneCorr [%]** - procento správně klasifikovaných fonémů po dekódování

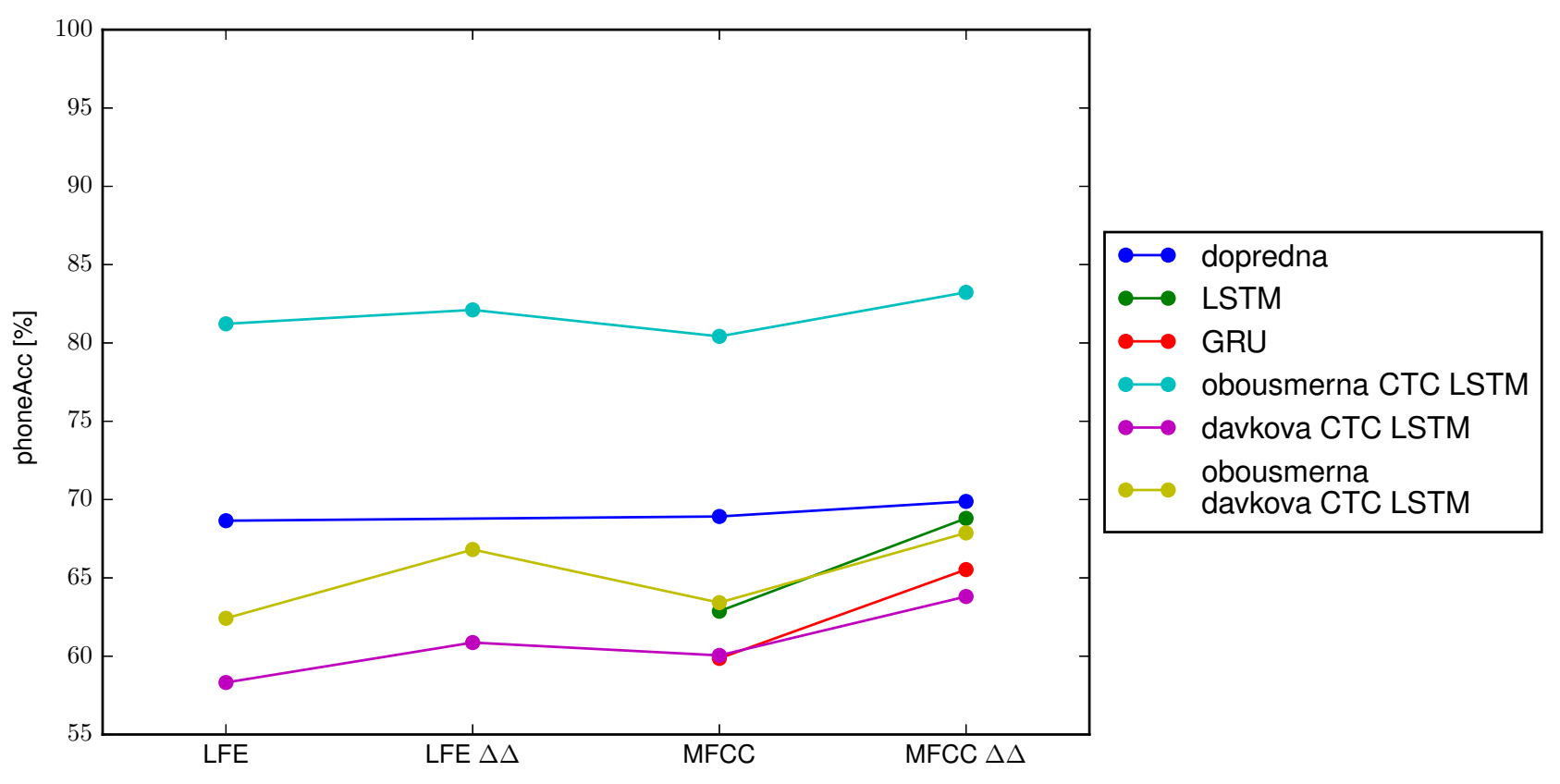
# Vyhodnocení

| | ŠkodaAuto | | | | SpeechDat-E | | | |
|---|---|---|---|---|---|---|---|---|
| architektura | LFE | LFE ΔΔ | MFCC | MFCC ΔΔ | LFE | LFE ΔΔ | MFCC | MFCC ΔΔ |
| dopředná | 79.99 | 80.15 | 79.63 | 81.06 | 68.65 | - | 68.92 | 69.88 |
| LSTM | 76.45 | 78.17 | 67.10 | 79.56 | - | - | 62.87 | 68.80 |
| GRU | 72.84 | 73.49 | 73.12 | 77.65 | - | - | 59.86 | 65.53 |
| obousměrná CTC LSTM | 92.97 | 93.34 | 93.03 | 93.70 | 81.22 | 82.11 | 80.42 | 83.23 |
| dávková CTC LSTM | 63.04 | 68.58 | 65.28 | 70.80 | 58.32 | 60.87 | 60.05 | 63.81 |
| obousměrná dávková CTC LSTM | 72.31 | 72.74 | 72.19 | 75.19 | 62.42 | 66.81 | 63.42 | 67.87 |

# Vyhodnocení - ŠkodaAuto



# Vyhodnocení - SpeechDat-E



# Shrnutí

- navrženo a porovnáno šest architektur neuronových sítí
- akcelerační koeficienty zvyšují přesnost rozpoznávání
- nejvyšší přesnost rozpoznání pro obousměrnou LSTM síť s CTC
  - přes 90% na datové sadě ŠkodaAuto
  - přes 80% na datové sadě SpeechDat-E
- rozpoznávání na základě znalosti celé nahrávky $\rightarrow$ nevhodné pro rozpoznávání v reálné čase
- možná rozšíření práce:
  - optimalizace topologie LSTM nebo dávkové obousměrné LSTM sítě s CTC
  - nalezení kompromisu mezi velikostí sítě, délkou vstupní sekvence a dostatečnou přesností pro rozpoznávání v reálném čase