Voice commands recognition in intelligent systems using deep neural networks

National Research University Higher School of Economics

Artem Sokolov, Andrey V. Savchenko

ASR classic pipeline. Problems

Acoustic model output:

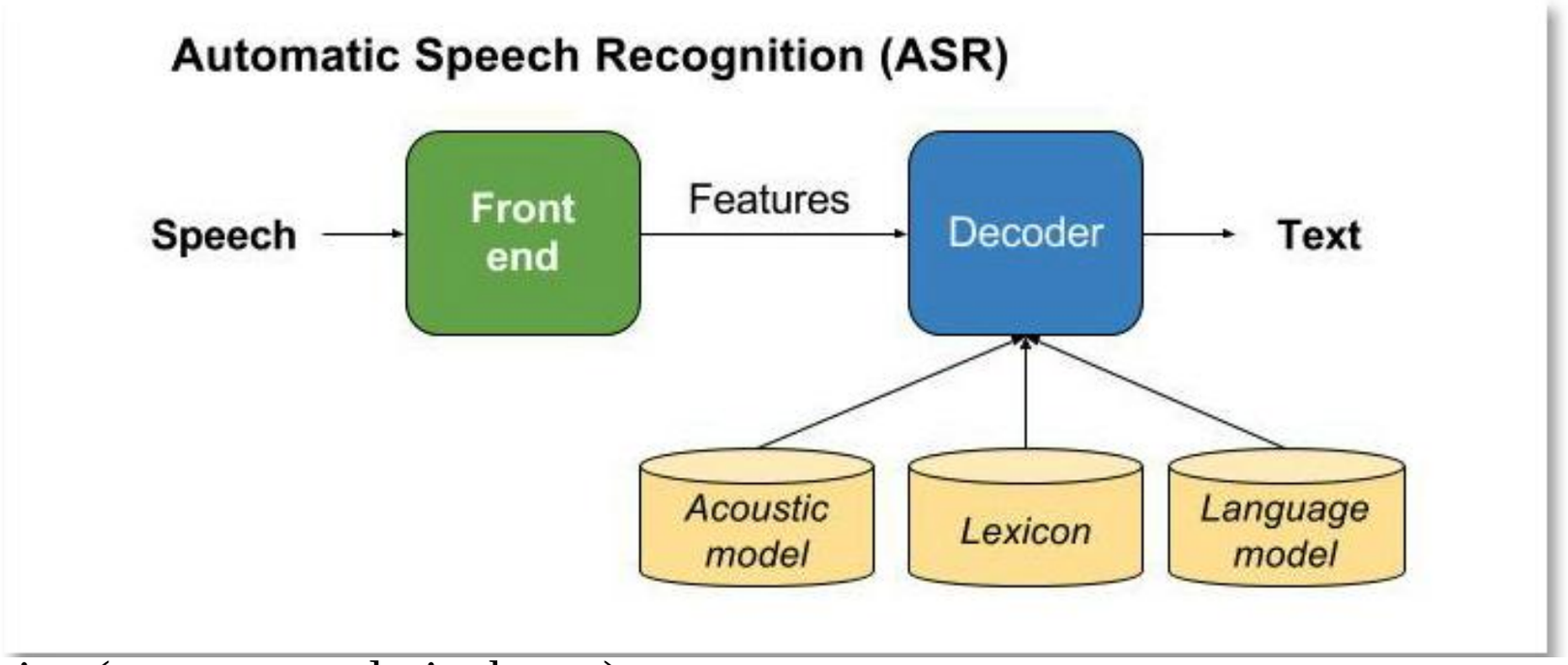
kaet

Lexicon:

k ae t→cat

Language model:

the cat→the cat



- Language model has a large size (up to several gigabytes)
- Significant computational resources required for quick inference. (ex. 2-3 word decoding takes tens of seconds on embedded device)
- Low accuracy with distorted data
- Accuracy depends on speaker pronunciation
- Solutions with the best accuracy required the Internet access.

Image credit: https://www.techrepublic.com/article/how-we-learned-to-talk-to-computers/

Goals and Tasks

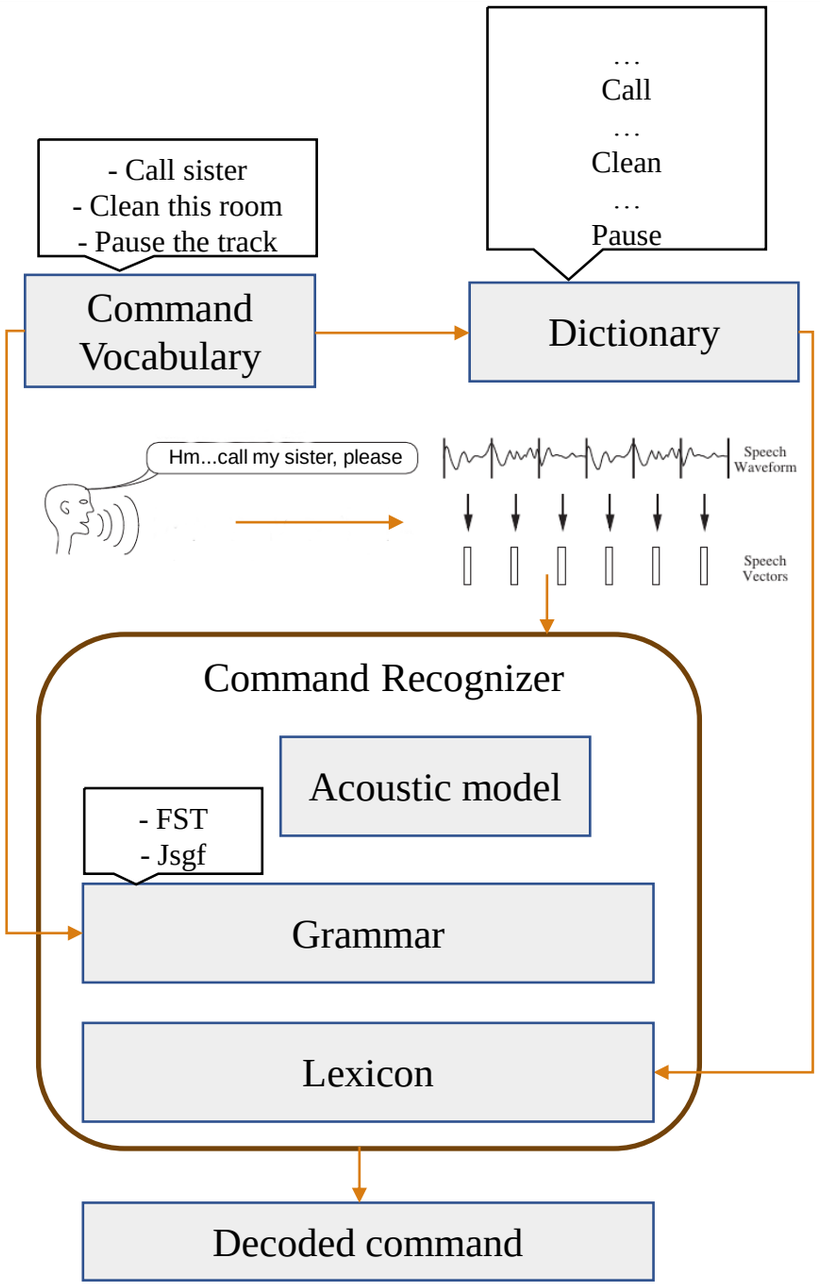
Goals:

- To Create a grammar-based ASR with accurate transcription for an abstract robotic system with a limited set of commands that works in noisy environment

Tasks:

- Create small command vocabulary (70-100 commands)
- Record commands and distort files
- Create Grammar based model
- Compare different offline frameworks with Grammar and Language models

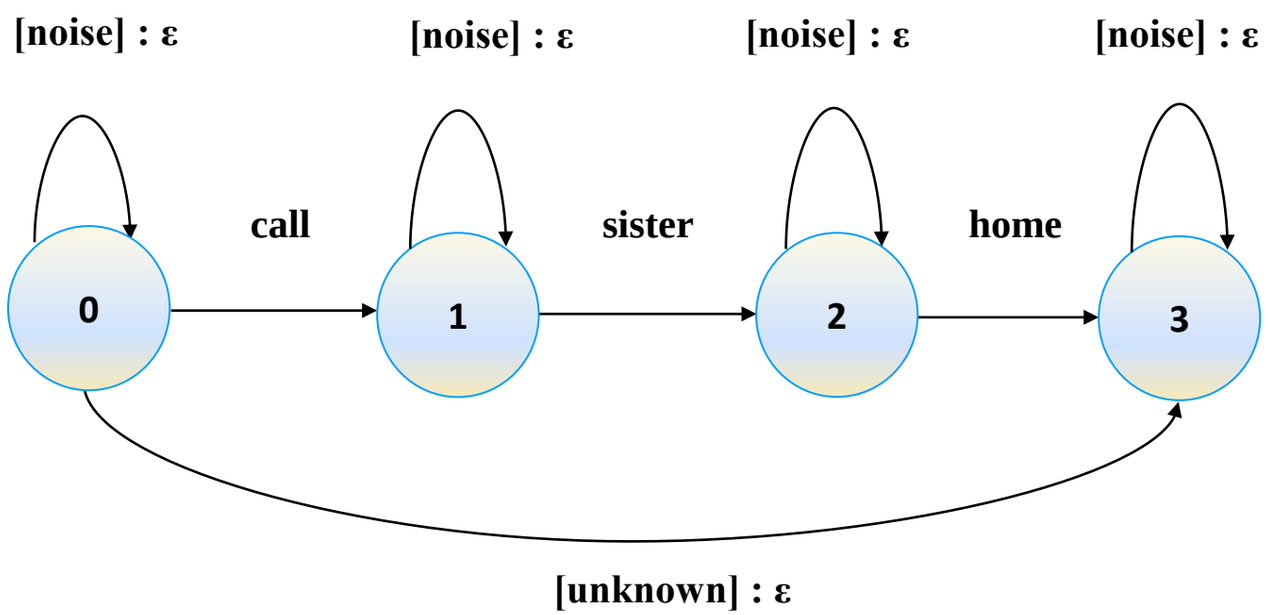
Our approach



Frameworks (used formats):

- Kaldi (FST, WFST)
- PocketSphinx (jsgf, arpa)
- Mozilla Deep Speech (arpa)
- Google speech API*

Grammar model example



The branch of a grammar graph

Data

- Created dictionary (for lexicon) includes 80 words

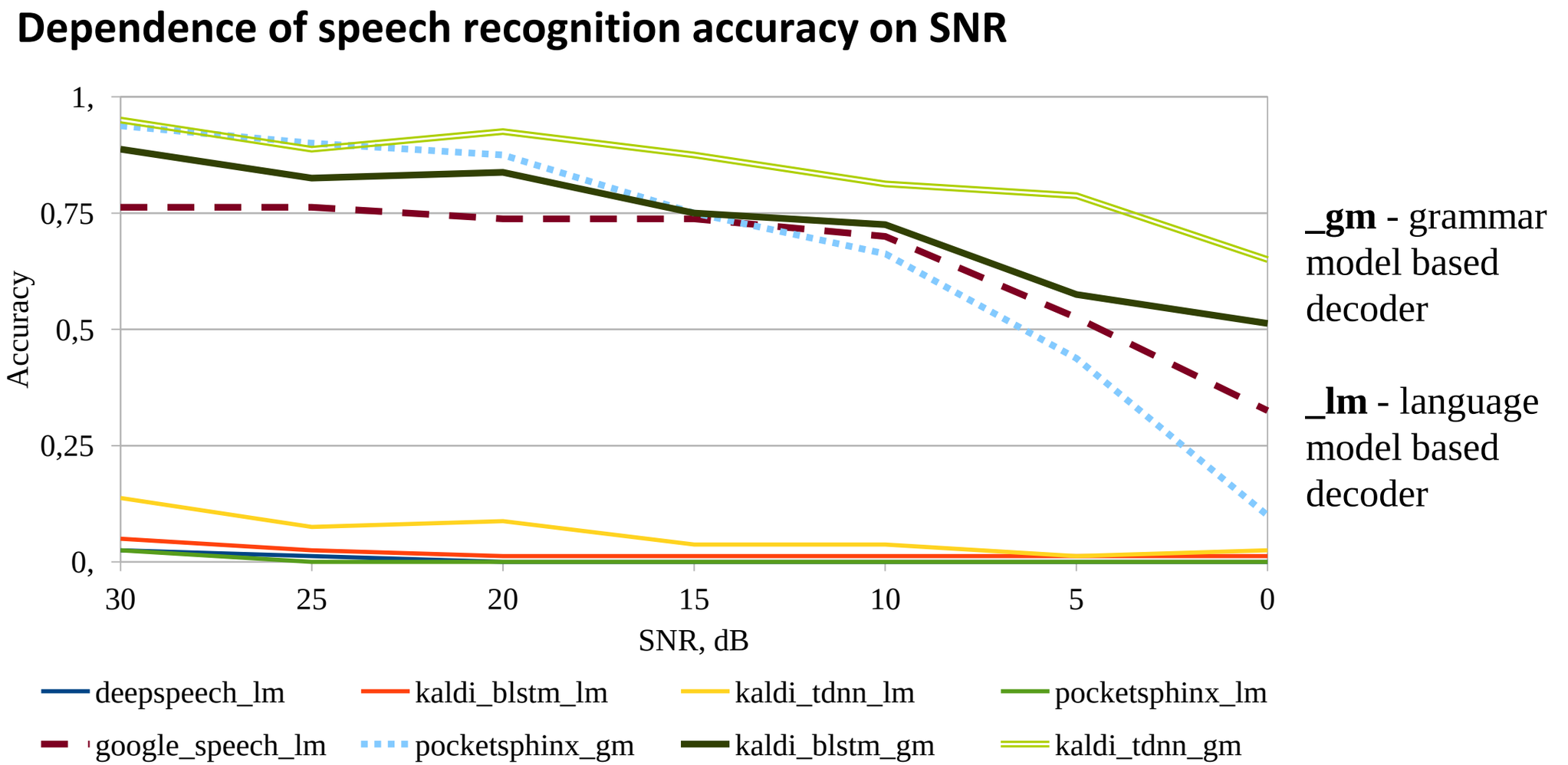
- Recorded 80 command (created grammar supports ~110) with non native speaker male voice and 40 with female voice.

- Files distorted with Sound Noise Ratio = 30(original), 25, 20, 15, 10, 5, 0

Compared approaches

- deepspeech_**lm** – the deepspeech implementation from mozilla
- pocketsphinx_**lm** – gmm based speech recognition framework with language model
- pocketsphinx_**gm** - gmm based speech recognition engine with grammar.
- google_speech_**lm** – the api speech toolkit from Google
- kaldi_blstm_**gm** – the kaldi speech recognition framework. (Used pretrained ASPIRE model based on blstm layer with grammar).
- kaldi_blstm_**lm** – the kaldi speech recognition framework. (Used pretrained ASPIRE model based on blstm layer with language model).
- kaldi_tdnn_**gm** – the kaldi speech recognition framework. (Used pretrained ASPIRE model based on tdnn layer with grammar).
- kaldi_tdnn_**lm** – the kaldi speech recognition framework. (Used pretrained ASPIRE model based on blstm layer with language model).

Experimental Results (I).

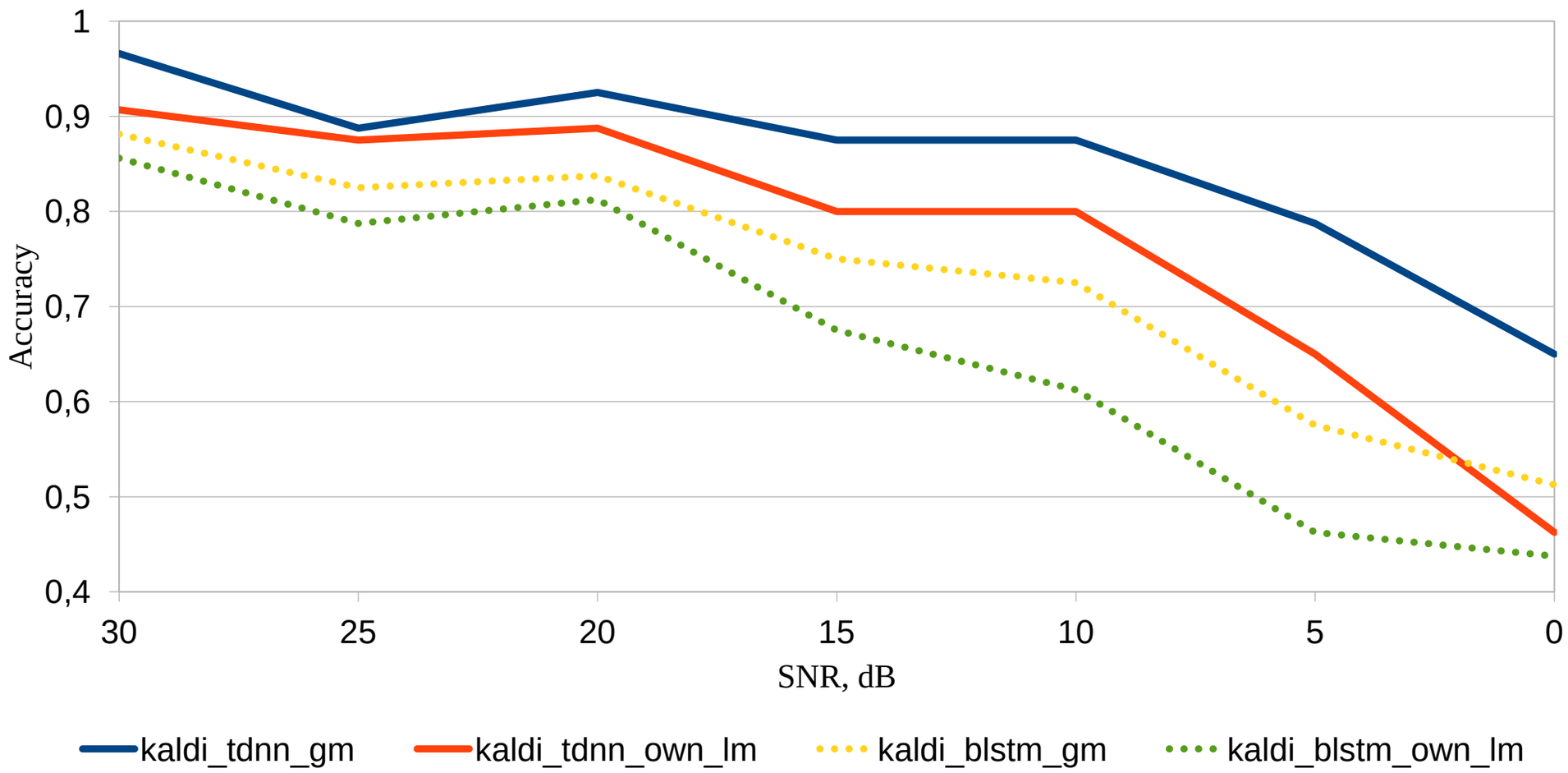


_gm - grammar model based decoder

_lm - language model based decoder

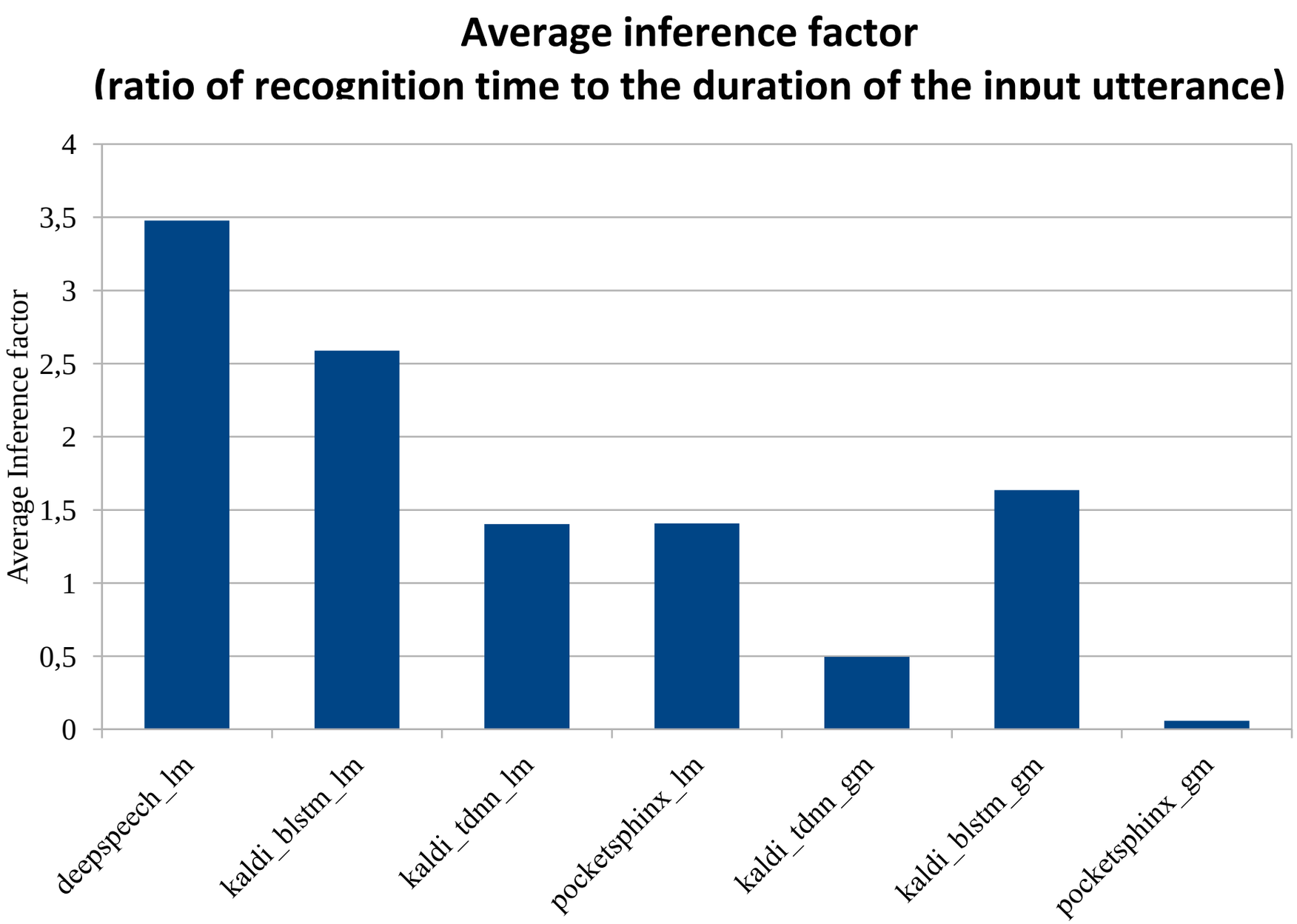
- GM-based approach obtains 60% higher accuracy when compared to the same models with LM.
- It is up to 20% higher than the proprietary Google Speech Recognition API.
- GMs have higher resistance to variations of noise and pronunciation

Experimental Results (II).



The language models trained on command corpus

Experimental Results (III). The speed of the inference.



Conclusion and future work

- Created Grammar Model for command vocabulary with up to 110 various commands
- Showed that such model has better accuracy and inference time than other LM based offline models.

Plans:

- Train own Seq2Seq based model on free data with TF framework
- Train gender-specific models and automatically choose the gender from facial video

Questions?