Comparing detection methods for pause-internal particles (PINTs)

Mikey Elmers

Department Language Science and Technology at Saarland University

elmers@lst.uni-saarland.de

IDeaL SFB 1102

DFG Deutsche Forschungsgemeinschaft

UNIVERSITÄT DES SAARLANDES

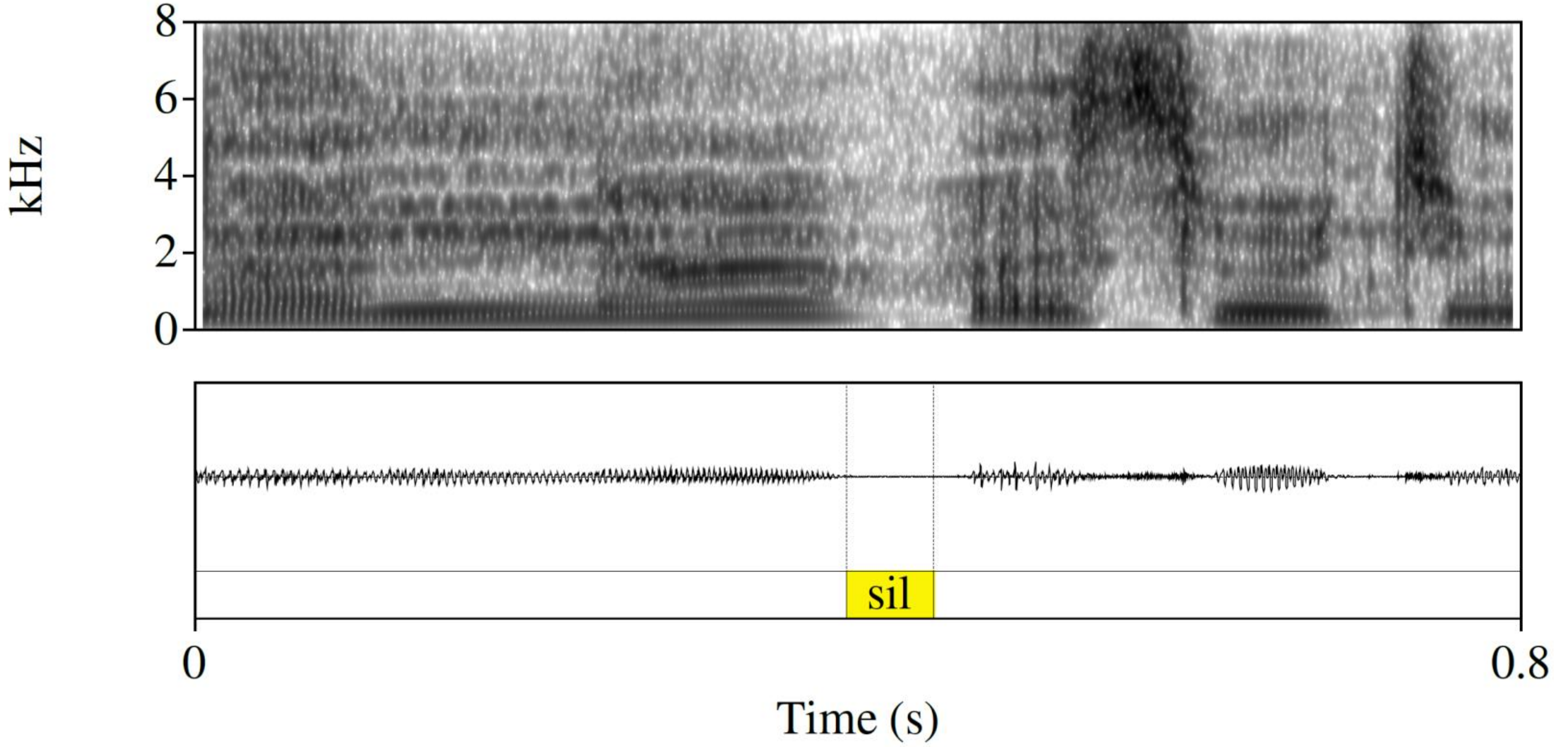
Introduction

- Silent segments
- Breath noises
  - Inhalations
  - Exhalations
- Filler particles
  - „äh“ and „ähm“ in German
  - „uh“ and „uhm“ in English
- Tongue clicks

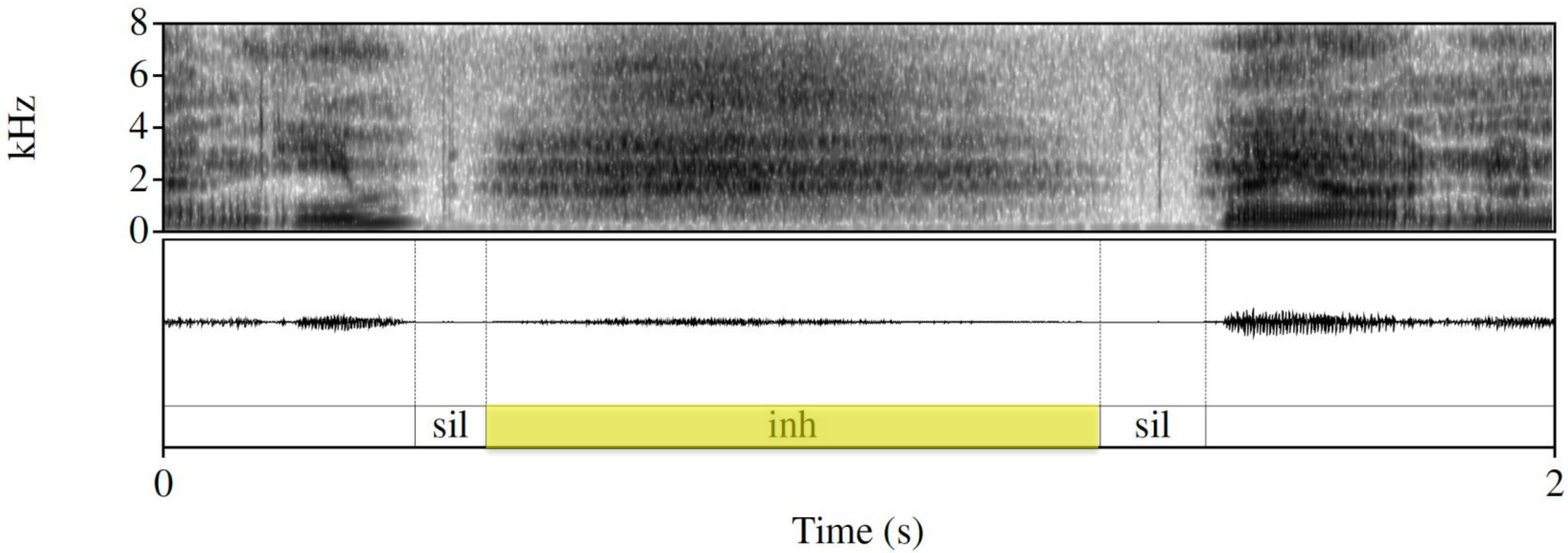
PINTs TTS

- Silent segments improve digit recollection (Elmers et al. 2021a)
- Breath noises improve sentence recollection (Elmers et al. 2021b)
- Filler particles improve TTS by reducing cognitive load for listener (Dall et al. 2016)
- Quality of training data is important for TTS applications (Henter et al. 2016)

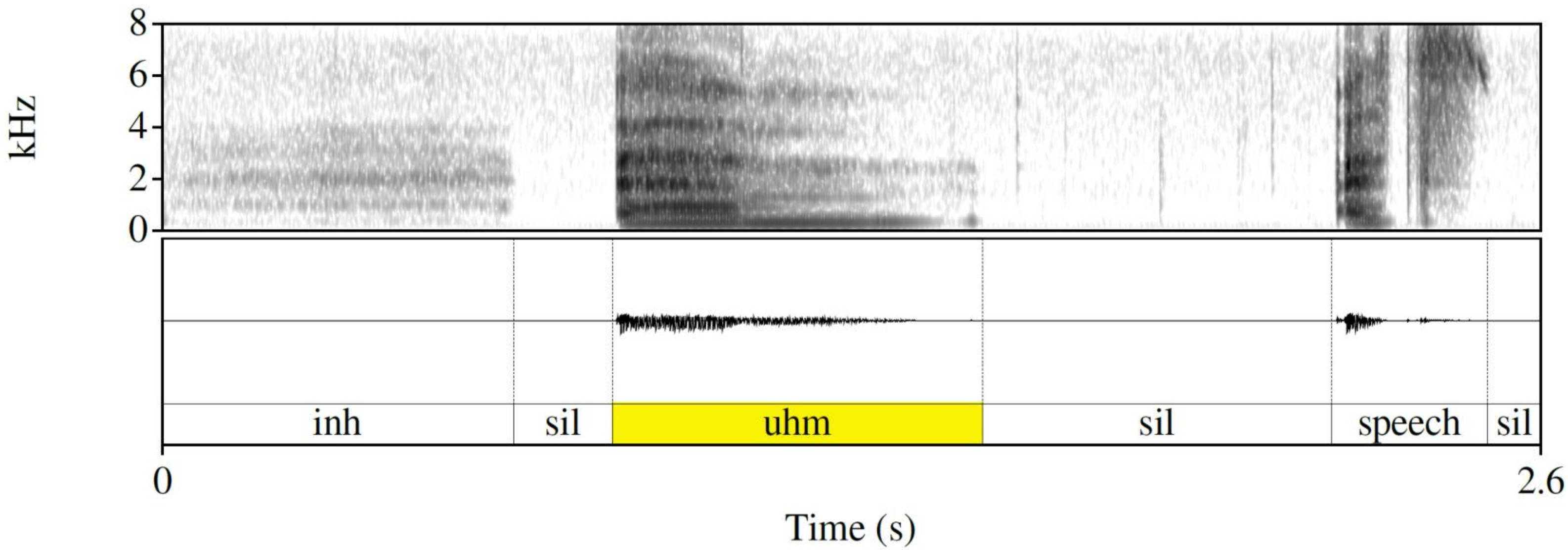
Silent Segment



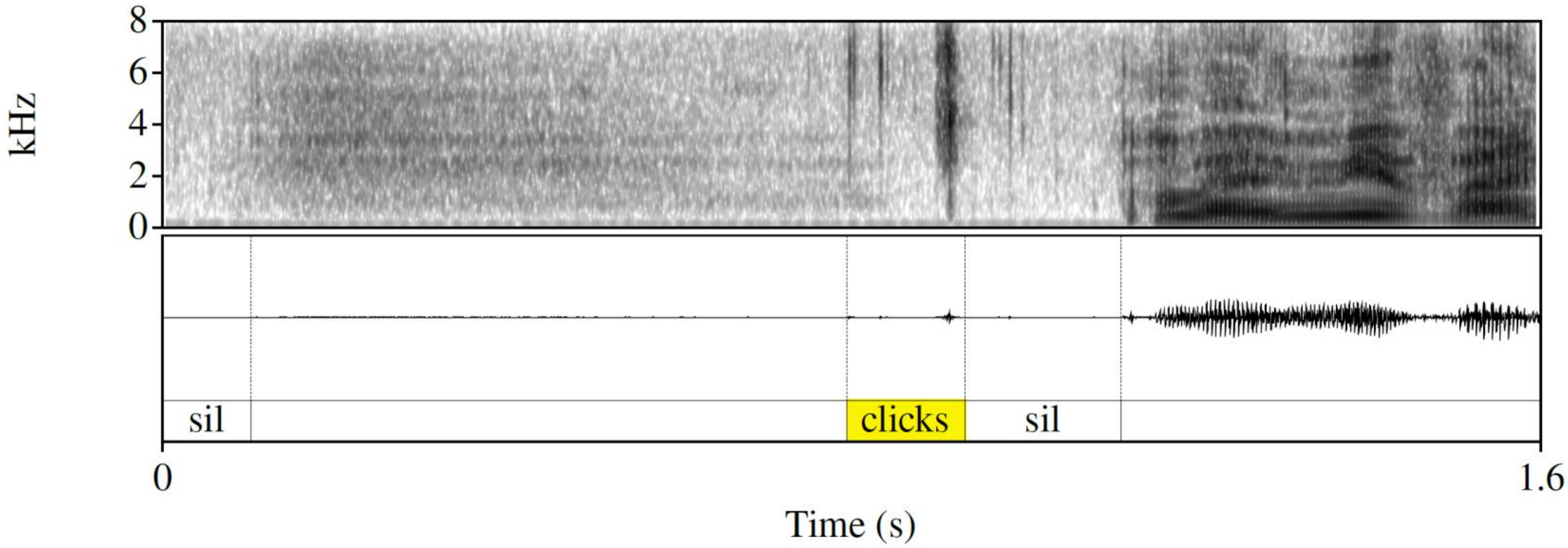
Breath Noises



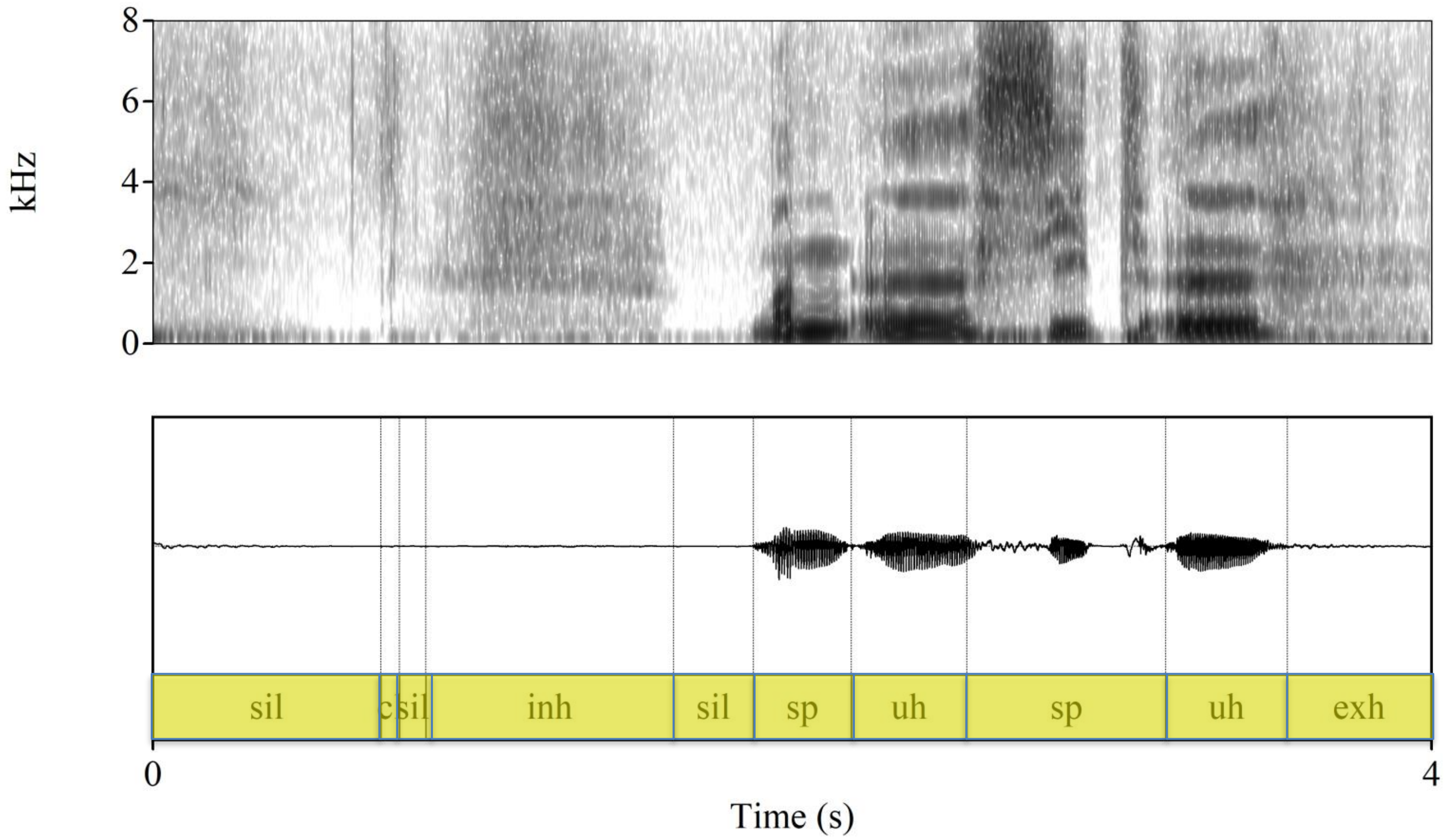
Filler Particles



Clicks



Co-Occurrence



Co-Occurrence

- Modeling multiple PINTs improved classification accuracy of surrounding non-verbal vocalizations (Condron et al. 2021)

- PINTs are usually:
  - Condensed to “other” class
  - Ignored altogether

Aim

- Implement state-of-the-art methods for detecting PINTs
- Classification of PINTs in German
- Classify PINTs using three models:
  - General neural network (NN)
  - Convolutional neural network (CNN)
  - Recurrent neural network (RNN)
- Hypotheses:
  - RNN will outperform other models
  - Simultaneous modeling improves PINTs classification

Methods

- Corpus Information:
  - Pool Corpus (Jessen et al. 2005)
  - 100 males (21-63 years old; mean age 39 years old)
  - Native speakers of German
  - Spontaneous speech task (i.e. picture description task)
  - Similar to board game Taboo

Methods

- Annotations:
  - 100 files (124-374 s; mean dur 223 s; total dur 6.2 hours)
  - Sampled at 16 kHz
  - 17,641 annotated PINTs
    - Silent segments, inhalations, exhalations, two types of filler particles („uh“ and „uhm“), and clicks
  - Other PINTs and disfluencies were excluded due to their infrequent occurrence

Methods

- Annotated PINTs overview
  - Min, max, mean, and sd measured in seconds
  - Total measured in minutes

class	count	min	max	mean	sd	total	prop
silent segment	10,237	0.01	20.01	0.65	0.95	111.04	29.92%
inhalation	2,891	0.05	2.10	0.51	0.27	24.79	6.68%
exhalation	1,887	0.03	3.23	0.38	0.28	12.15	3.27%
filler (uh)	1,156	0.04	1.44	0.35	0.16	6.81	1.83%
filler (uhm)	549	0.15	2.64	0.53	0.25	4.85	1.30%
click	921	0.00	0.50	0.06	0.05	0.96	0.25%

Methods

- Data pre-processing:
  - 13 mel-frequency cepstral coefficients (MFCCs)
  - Frame size 93 ms
  - Hop length 23 ms
  - Zero-padding

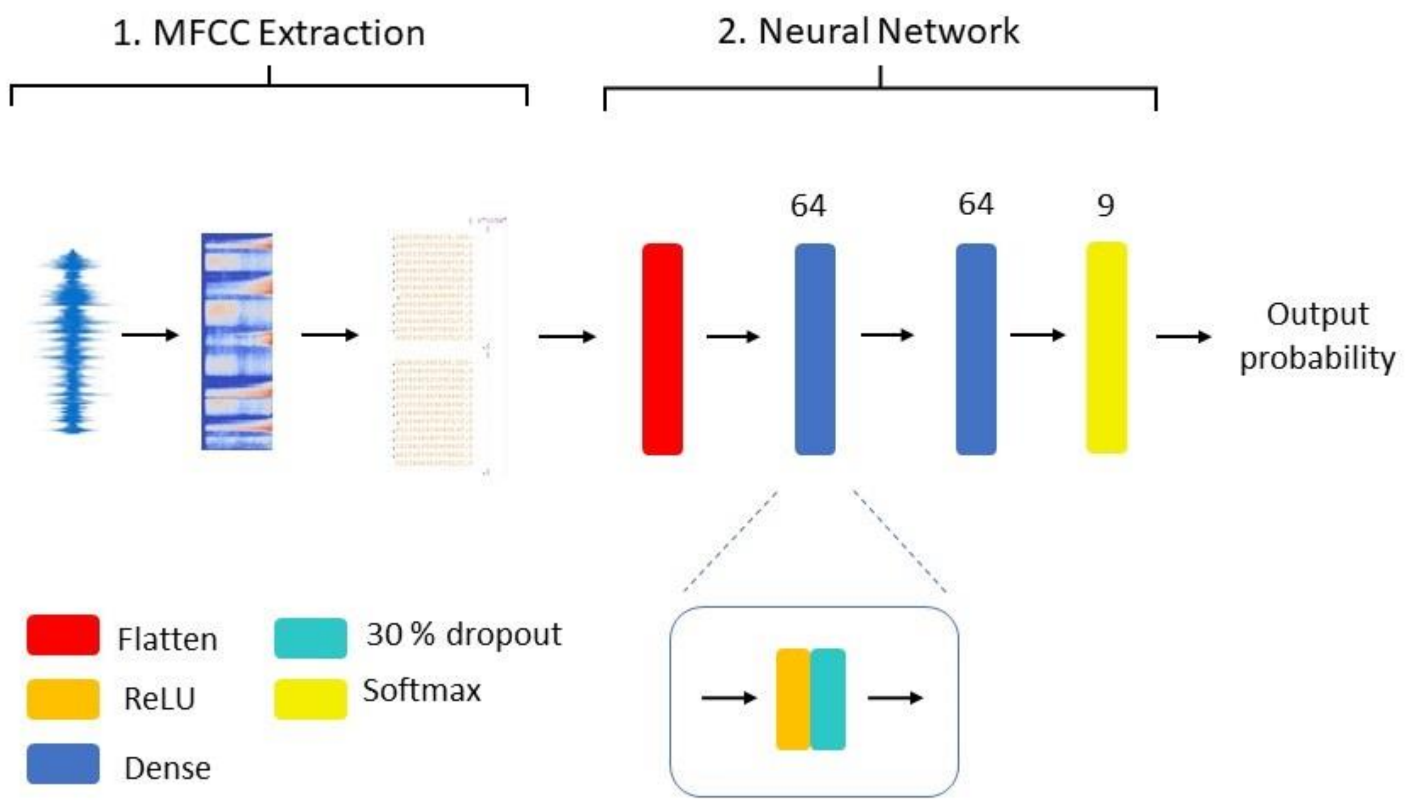
Methods

- Data pre-processing:
  - Models trained on nine classes
    - **Silent segments**
    - **Inhalation**
    - **Exhalation**
    - **Two FPs (“uh” and “uhm”)**
    - **Clicks**
    - Speech
    - Task change
    - Zero-padding

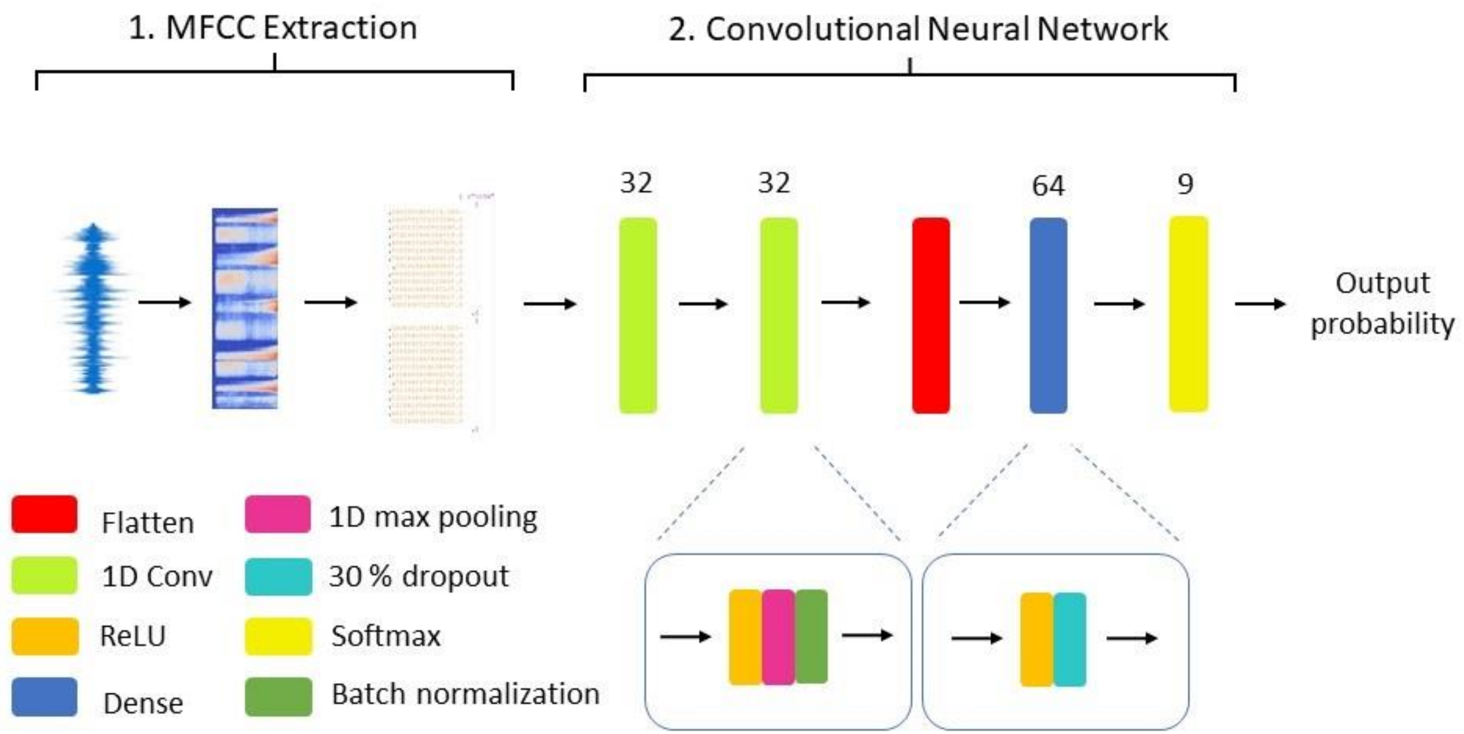
Methods

- Model Information:
  - Same hyperparameters
  - Similar number of layers
  - Same number of neurons for those layers
  - Sparse categorical cross entropy loss function
  - Learning rate of 0.0001
  - Adam optimizer
  - Batch size of 32
  - Trained for 40 epochs

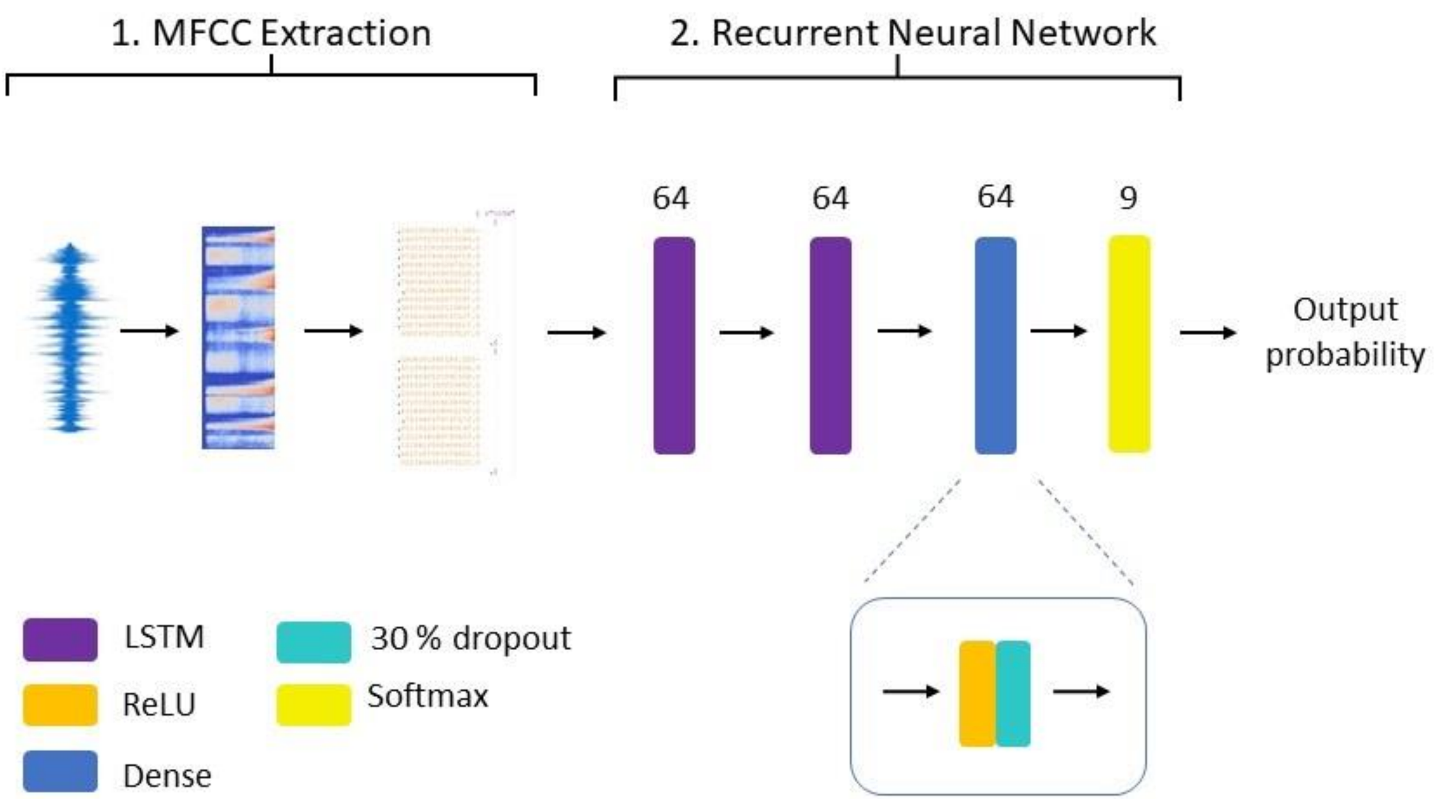
Methods – Neural Network



Methods – Convolutional Neural Network



Methods – Recurrent Neural Network



Results

NN							
class	**sil**	**inh**	**exh**	**uh**	**uhm**	**click**	**sum**
silent segment (sil)	64,971	2,743	789	-	-	-	68,503
inhalation	4,141	10,372	58	-	-	-	14,571
exhalation	3,215	497	2,188	-	-	-	5,900
filler (uh)	60	3	34	-	-	-	97
filler (uhm)	68	4	33	-	-	-	105
click	209	85	6	-	-	1	301
sum	72,664	13,704	3,108	-	-	1	89,477

CNN							
class	**sil**	**inh**	**exh**	**uh**	**uhm**	**click**	**sum**
silent segment (sil)	66,494	1,375	754	-	-	1	68,624
inhalation	5,111	9,351	100	-	-	-	14,562
exhalation	3,173	336	2,532	-	-	-	6,041
filler (uh)	53	2	27	-	-	-	82
filler (uhm)	80	5	20	-	11	-	116
click	181	73	11	-	-	-	265
sum	75,092	11,142	3,444	-	11	1	89,690

RNN							
class	**sil**	**inh**	**exh**	**uh**	**uhm**	**click**	**sum**
silent segment (sil)	64,771	1,813	811	-	-	-	67,395
inhalation	4,214	10,098	113	-	-	-	14,425
exhalation	2,812	394	2,308	-	-	-	5,514
filler (uh)	38	2	13	-	-	-	53
filler (uhm)	50	2	17	-	3	-	72
click	165	74	8	-	-	3	250
sum	72,050	12,383	3,270	-	3	3	87,709

Results

Model	Accuracy	Precision	Recall	F1 Score
NN	85.6%	53.5%	41.6%	40.5%
CNN	86.1%	53.2%	41.9%	41.8%
RNN	86.1%	69.0%	42.1%	41.7%

Model	sil	inh	exh	uh	uhm	click
NN	94.8%	71.2%	31.1%	0.0%	0.0%	0.3%
CNN	96.9%	64.2%	41.9%	0.0%	9.5%	0.0%
RNN	96.1%	70.0%	41.9%	0.0%	4.2%	1.2%

Conclusions

- All models performed similarly
- Hypotheses:
  - 1) RNN should perform best since it considers temporal information
    - RNN did not perform much better than NN or CNN

Conclusions

- Hypotheses:
  - 2) Simultaneous modeling can improve classification accuracy of surrounding PINTs
    - Simultaneous modeling didn‘t improve accuracy for surrounding PINTs
    - All models unable to classify FPs and clicks
    - FPs too close to speech category
    - Clicks often misclassified as silent segments
      - short duration
      - drawback of only using MFCCs as input

Conclusions

- Model classified:
  - Silent segments very well
  - Inhalations well
  - Exhalations with middling success

- Accurate PINTs classification dependent on:
  - Annotation quality
  - Annotation quantity
  - Models started with high accuracy and improved minimally

Conclusions

- Improvement to PINTs detection:
  - Increase number of occurrences
  - Especially for infrequent PINTs

- Future work
  - Investigate other acoustic features
  - Train using spectrogram images
  - Implement PINTs classification into TTS pipeline

Reference

- CONDRON, S., G. CLARKE, A. KLEMENTIEV, D. MORSE-KOPP, J. PARRY, and D. PALAZ: Non-verbal vocalisation and laughter detection using sequence-to-sequence models and multi-label training. In Proc. Interspeech 2021, pp. 2506–2510. 2021.

- DALL, R., M. TOMALIN, and M. WESTER: Synthesising filled pauses: Representation and datamixing. In 9th ISCA Speech Synthesis Workshop, pp. 7–13. 2016.

- ELMERS, M., R. WERNER, B. MUHLACK, B. MÖBIUS, and J. TROUVAIN: Evaluating the effect of pauses on number recollection in synthesized speech. In Elektronische Sprachsignalverarbeitung 2021, Tagungsband der 32. Konferenz, Studientexte zur Sprachkommunikation, pp. 289–295. TUD Press, Berlin, 2021.

- ELMERS, M., R. WERNER, B. MUHLACK, B. MÖBIUS, and J. TROUVAIN: Take a breath: Respiratory sounds improve recollection in synthetic speech. In Proc. Interspeech 2021, pp. 3196–3200. 2021.

- HENTER, G. E., S. RONANKI, O. WATTS, M. WESTER, Z. WU, and S. KING: Robust tts duration modelling using dnns. In 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5130–5134. IEEE, 2016.

- JESSEN, M., O. KÖSTER, and S. GFROERER: Influence of vocal effort on average and variability of fundamental frequency. International Journal of Speech Language and the Law, 12(2), pp. 174–213, 2005.

PINTS Website

Thank you!

http://pauseparticles.org/

Conclusions

- Model classified:
  - Silent segments very well
  - Inhalations well
  - Exhalations with middling success

- Accurate PINTs classification dependent on:
  - Annotation quality
  - Annotation quantity
  - Models started with high accuracy and improved minimally