Philipps Universität Marburg

GEFÖRDERT VOM

Bundesministerium für Bildung und Forschung

Finanziert von der Europäischen Union
NextGenerationEU

EDAudio:
Easy Data Augmentation Techniques for Audio Classification

REDE
regionalsprache.de

Akademie der Wissenschaften und der Literatur | Mainz

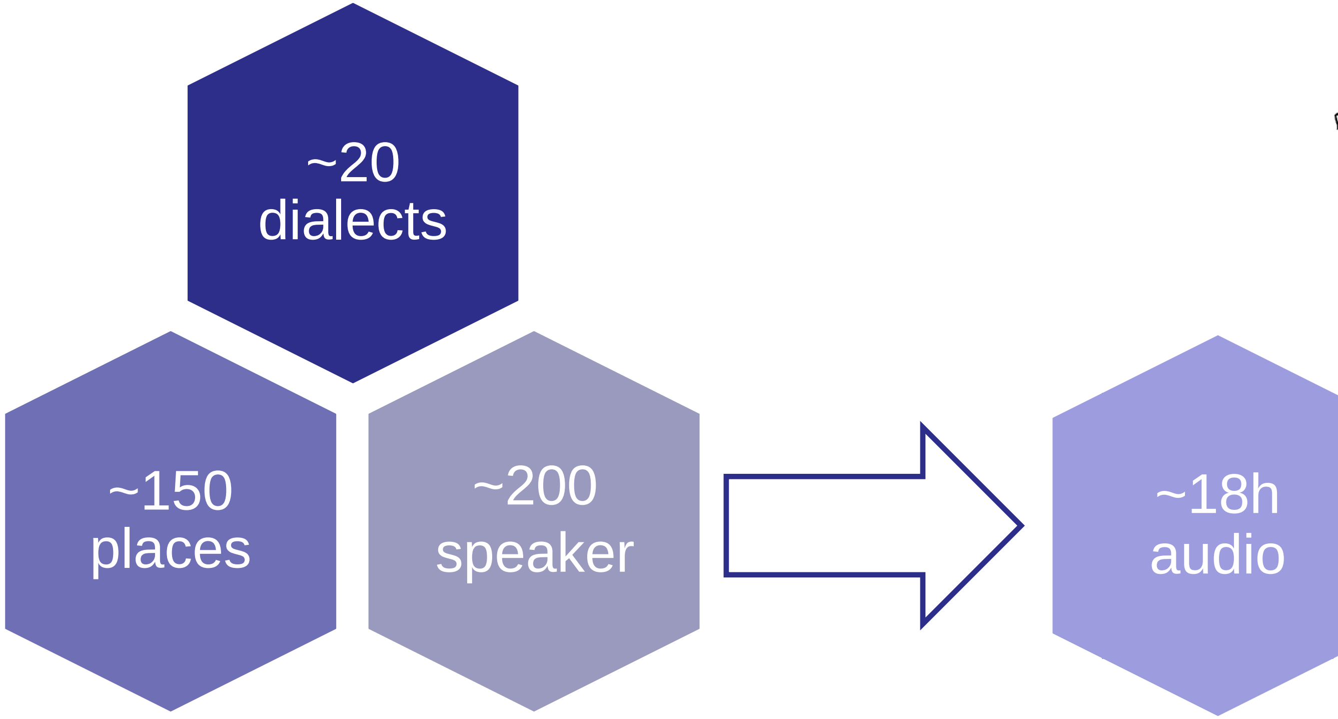
Alfred Lameli
Lea Fischbach
Caroline Kleen
Akbar Karimi
Lucie Flek

Forschungszentrum Deutscher Sprachatlas
DSA

10th July 2024, ICLaVE - Vienna

Overview

- Recordings from Regionalsprache.de (2008-2012)
- Used Speaker are male and +65years
- Used Recordings contain dialectal „Wenker phrases“[1]

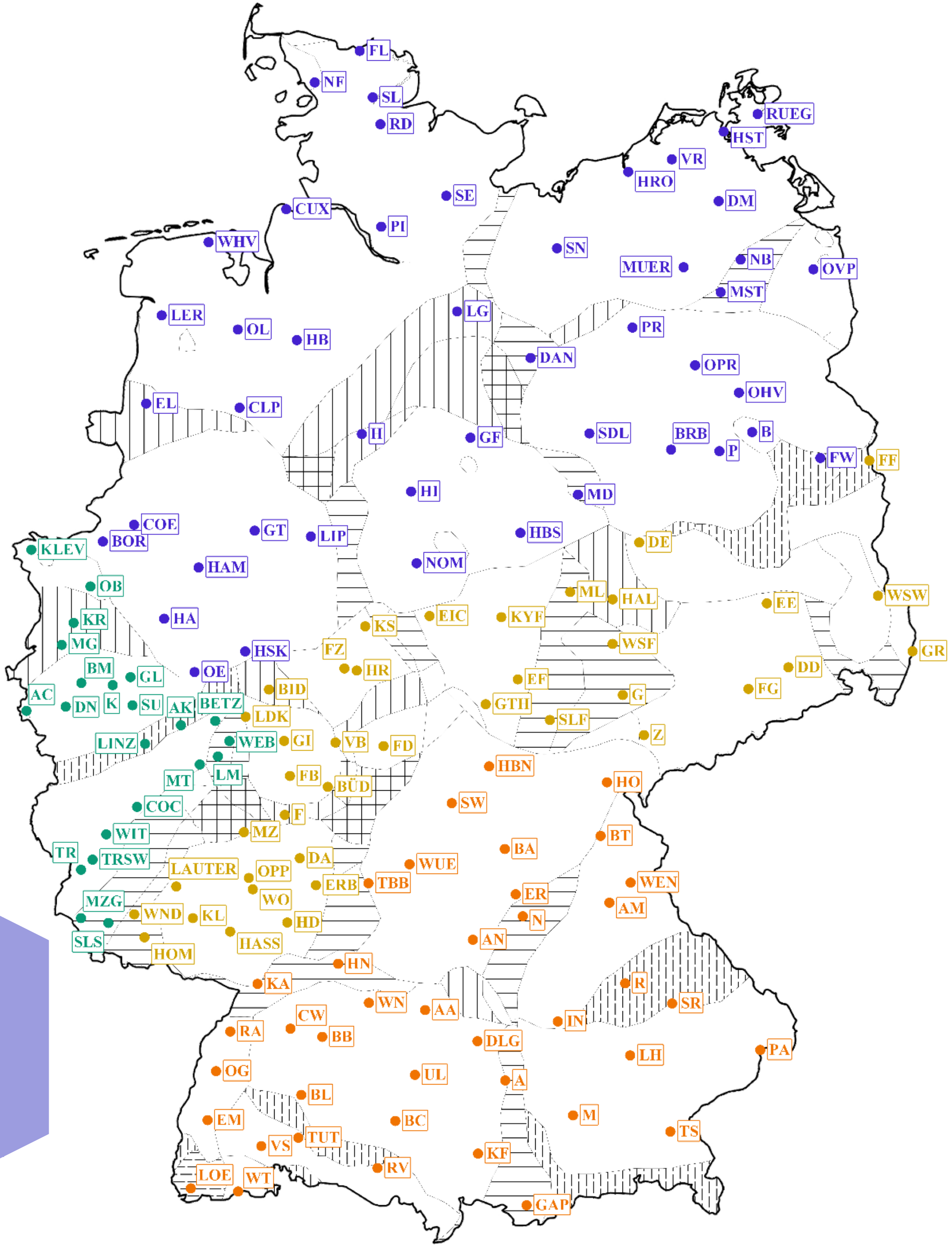


~20 dialects

~150 places

~200 speaker

~18h audio



[1] https://www.uni-marburg.de/en/fb09/dsa/research-documentation-center/wenkersaetze

Why Augmentation?

- Increase in Dataset Size
- Regularization
- Balancing Classes

⇨ Improvement in Model Performance

Experiment Motivation

Spectrogram Data Augmentation (SDA)

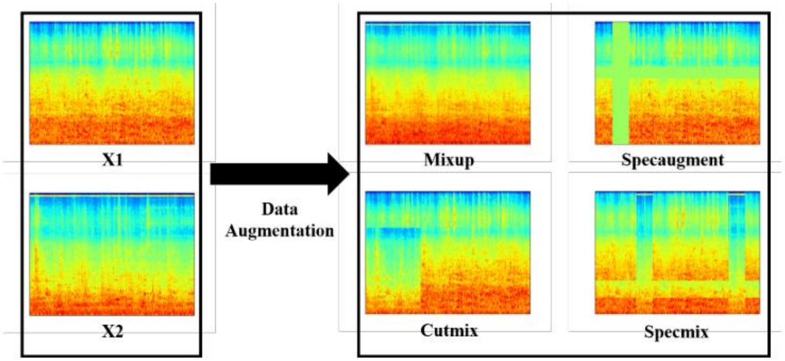


Figure 1: *Overview of the data augmentation methods:Mixup, Cutmix, Specaugment and our Specmix.*

G. Kim, D. K. Han, and H. Ko, “Specmix: A mixed sample data augmentation method for training withtime-frequency domain features,” arXiv preprint arXiv:2108.03020, 2021

D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, “Specaugment: A simple data augmentation method for automatic speech recognition,” arXiv preprint arXiv:1904.08779, 2019

Environmental Sound Classification

J. Salamon and J. P. Bello, “Deep convolutional neural networks and data augmentation for environmental sound classification,” IEEE Signal processing letters, vol. 24, no. 3, pp. 279–283, 2017.

- speed modification yields the most significant improvement
- noise addition contributes the least

Not true for dialect classification

Automatic Speech Recognition (ASR)

T. Fukuda, R. Fernandez, A. Rosenberg, S. Thomas, B. Ramabhadran, A. Sorin, and G. Kurata, “Data augmentation improves recognition of foreign accented speech.” in Interspeech, no. September, 2018, pp. 2409–2413.

- Pitch Shift only method leading to improvement across all classes
- BN in charge of the least improvement

Not true for dialect classification

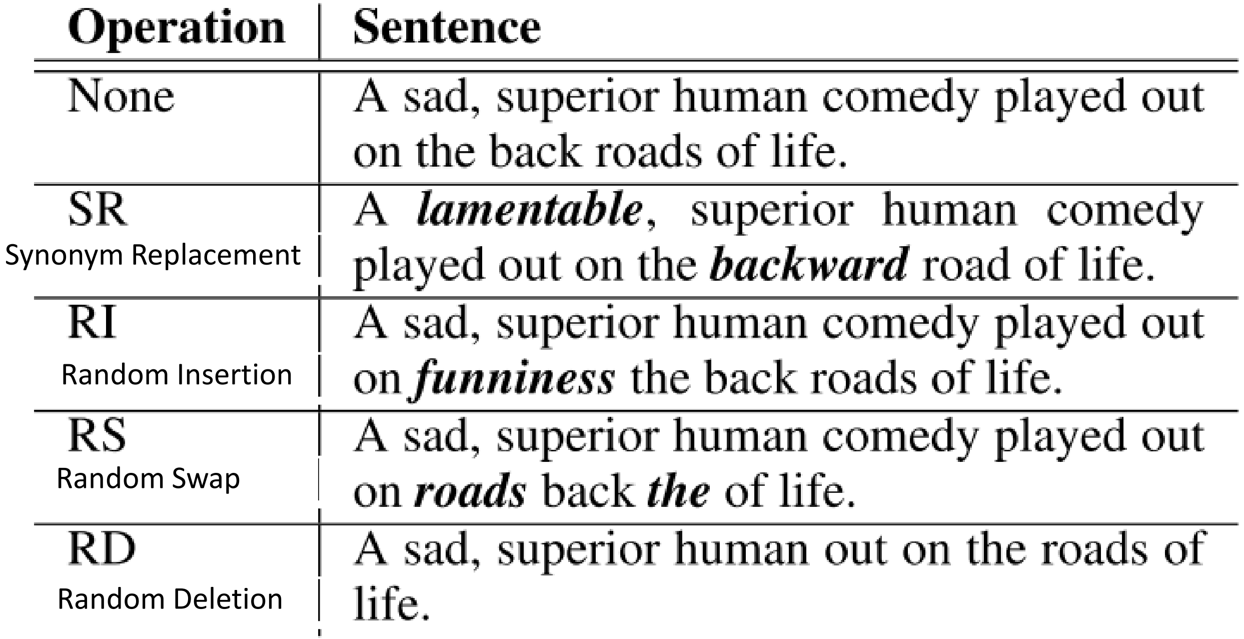
Experiment Setup

- Weighted f1-Score
- Cut Audio Files into 10-second Segments
- Fixed Speaker for training/validation/testing
- Run Model 50 times → get mean Score
- Starting with weighted f1_Score of 0.221

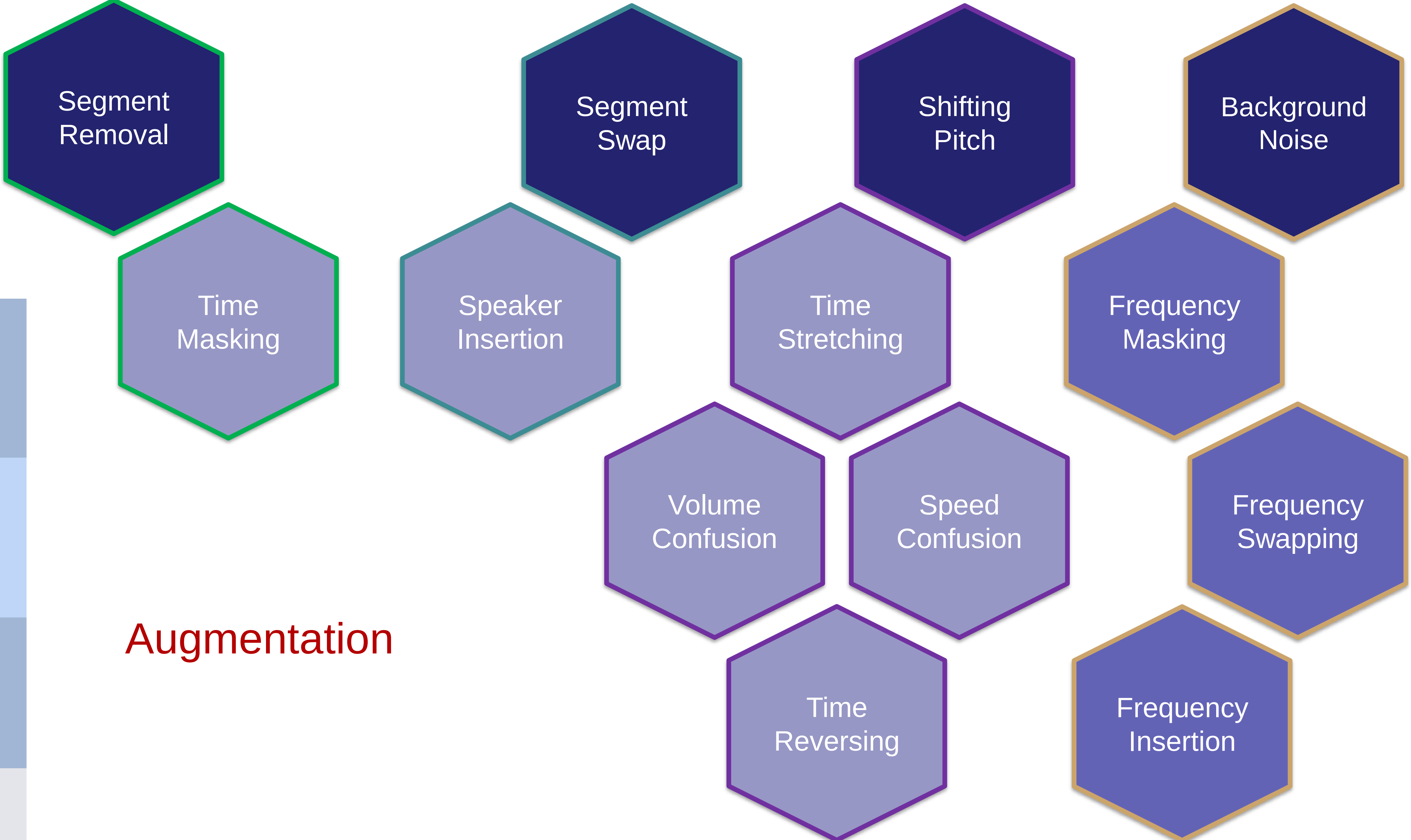
Experiment Motivation

Text Classification

J. Wei and K. Zou, “Eda: Easy data augmentation techniques for boosting performance on text classification tasks,” arXiv preprint arXiv:1901.11196, 2019



Operation	Sentence
None	A sad, superior human comedy played out on the back roads of life.
SR Synonym Replacement	A ***lamentable***, superior human comedy played out on the ***backward*** road of life.
RI Random Insertion	A sad, superior human comedy played out on ***funniness*** the back roads of life.
RS Random Swap	A sad, superior human comedy played out on ***roads*** back ***the*** of life.
RD Random Deletion	A sad, superior human out on the roads of life.



Segment Removal

Time Masking

Segment Swap

Speaker Insertion

Shifting Pitch

Time Stretching

Volume Confusion

Speed Confusion

Time Reversing

Background Noise

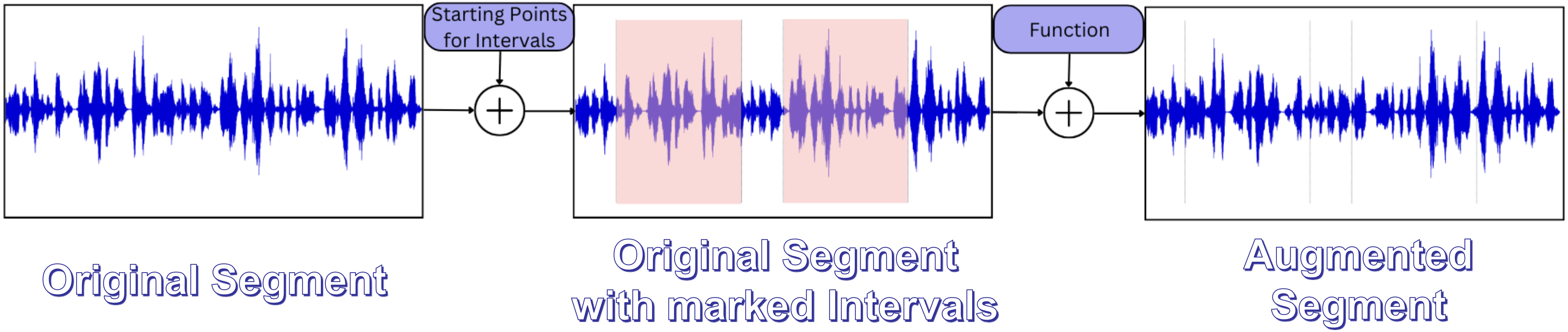
Frequency Masking

Frequency Swapping

Frequency Insertion

Augmentation

Experimental Setup


Starting Points for Intervals
Function
Original Segment
Original Segment with marked Intervals
Augmented Segment

Experimental Setup

Praat

P. Boersma and D. Weenink, “Praat: doing phonetics by computer [Computer program],” Version 6.1.38, retrieved 2 January 2021 http://www.praat.org/, 2021.

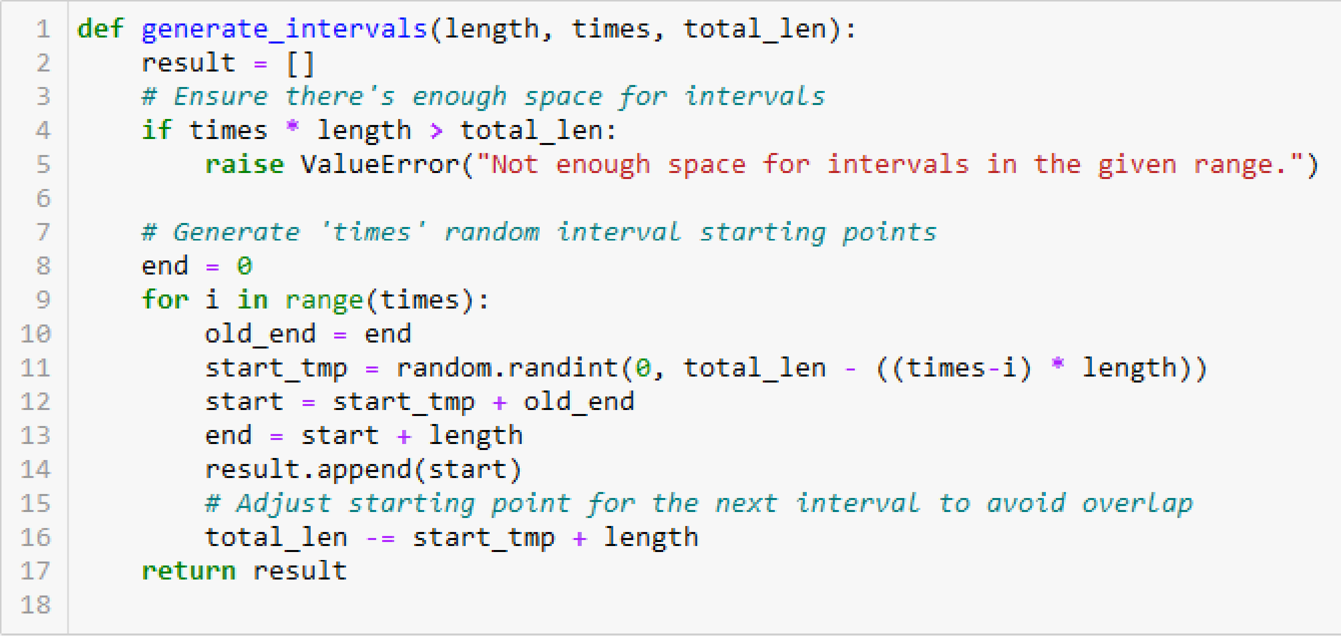
$$n_{aug} = (\propto * \, l_{audio})/l_{aug}$$

$\propto$

l_{aug}

	0.1	0.3	0.5	1.0
0.3	3	10	16	33
1	1	3	5	10
4	-	-	1	2
5	-	-	1	2
10	-	-	-	1

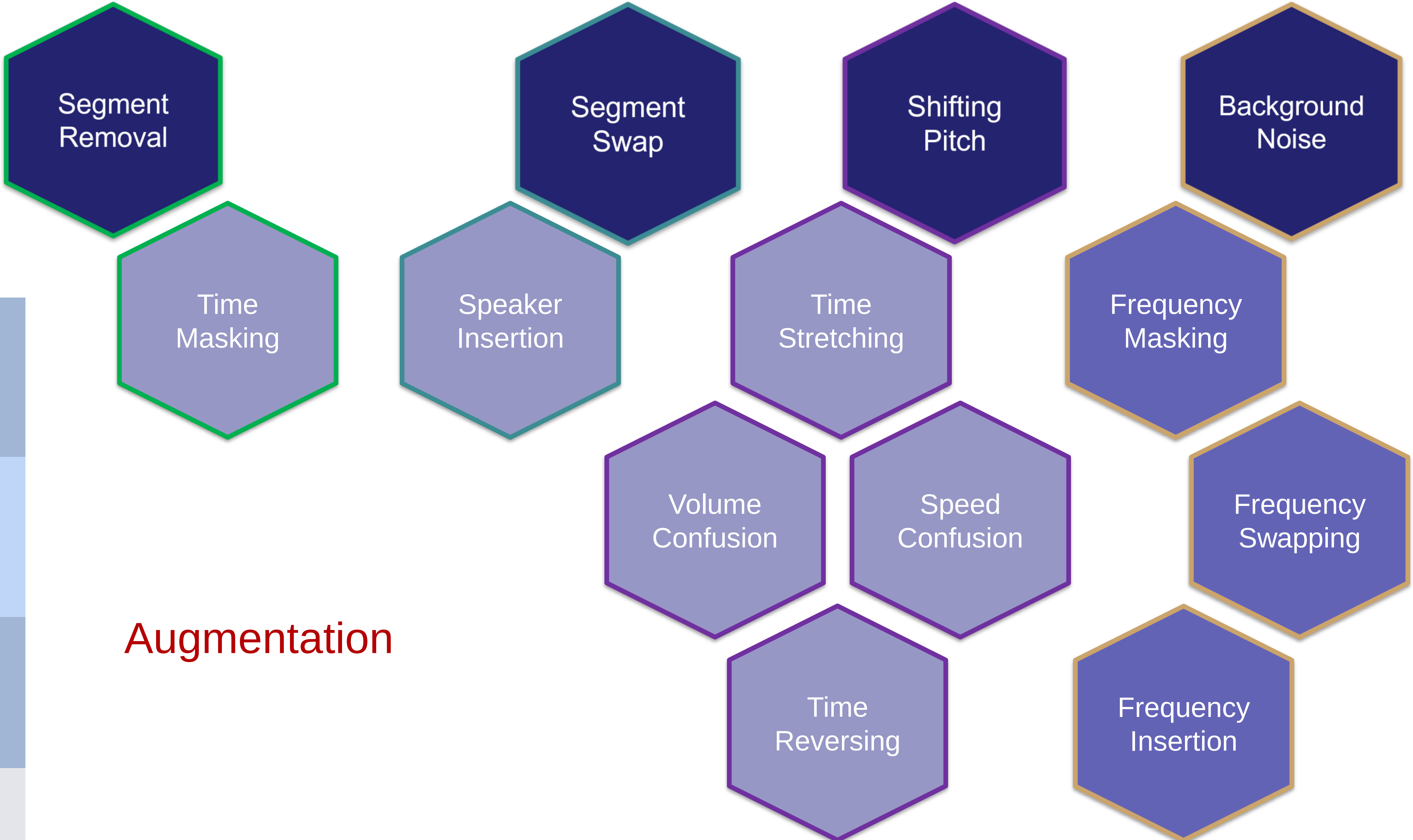
$$n_{augFiles} = \{1,2,4,6\}$$



```python
def generate_intervals(length, times, total_len):
    result = []
    # Ensure there's enough space for intervals
    if times * length > total_len:
        raise ValueError("Not enough space for intervals in the given range.")

    # Generate 'times' random interval starting points
    end = 0
    for i in range(times):
        old_end = end
        start_tmp = random.randint(0, total_len - ((times-i) * length))
        start = start_tmp + old_end
        end = start + length
        result.append(start)
        # Adjust starting point for the next interval to avoid overlap
        total_len -= start_tmp + length
    return result
```

Augmentation



Segment Removal

Time Masking

Segment Swap

Speaker Insertion

Shifting Pitch

Time Stretching

Volume Confusion

Speed Confusion

Time Reversing

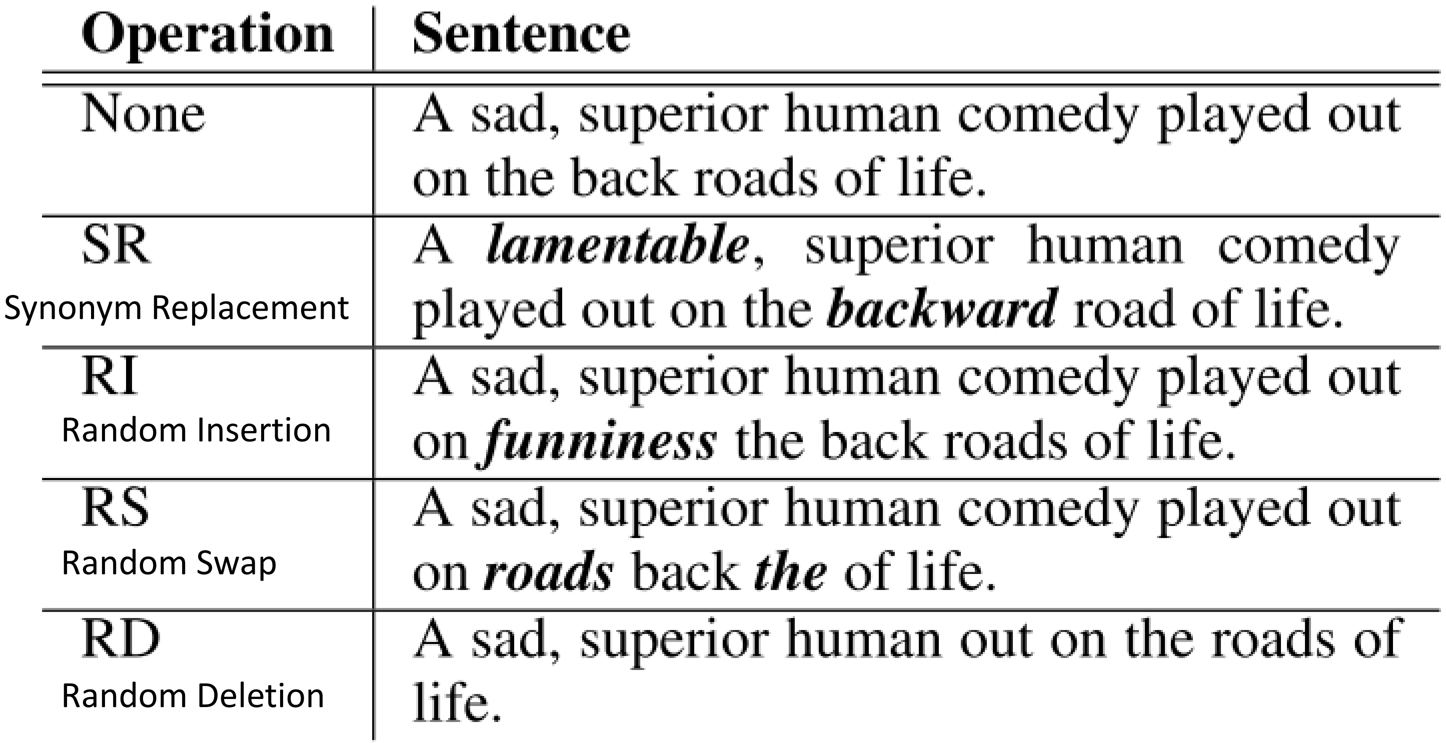
Background Noise

Frequency Masking

Frequency Swapping

Frequency Insertion

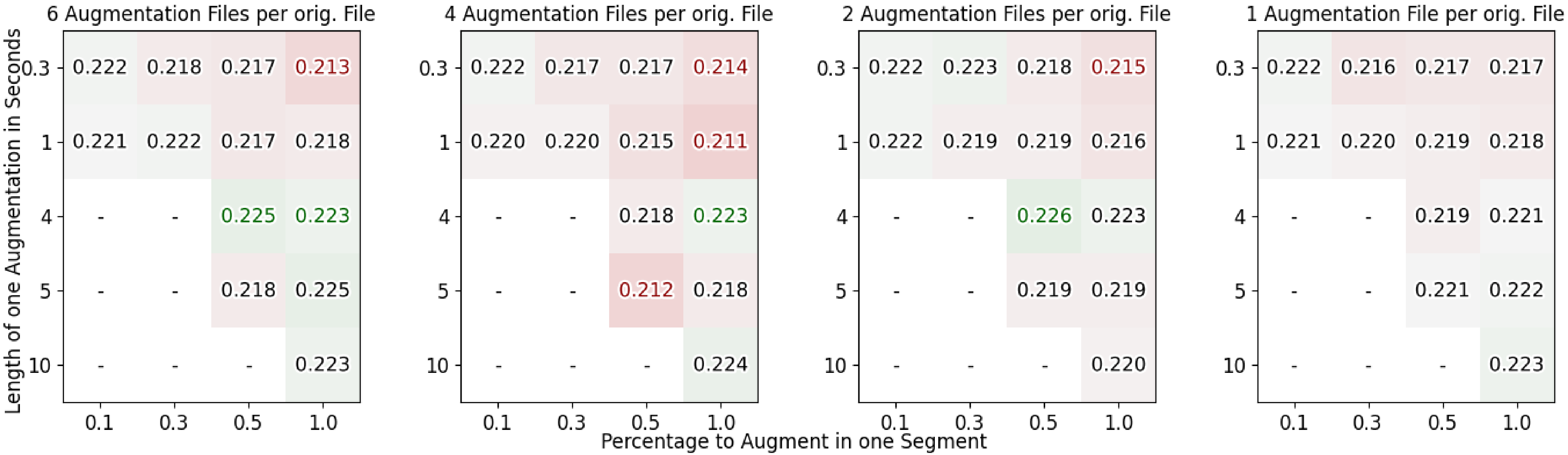
EDA



Operation	Sentence
None	A sad, superior human comedy played out on the back roads of life.
SR Synonym Replacement	A ***lamentable***, superior human comedy played out on the ***backward*** road of life.
RI Random Insertion	A sad, superior human comedy played out on ***funniness*** the back roads of life.
RS Random Swap	A sad, superior human comedy played out on ***roads*** back ***the*** of life.
RD Random Deletion	A sad, superior human out on the roads of life.

Shifting Pitch (SP)

Shifting Pitch



6 Augmentation Files per orig. File

	0.1	0.3	0.5	1.0
0.3	0.222	0.218	0.217	0.213
1	0.221	0.222	0.217	0.218
4	-	-	0.225	0.223
5	-	-	0.218	0.225
10	-	-	-	0.223

4 Augmentation Files per orig. File

	0.1	0.3	0.5	1.0
0.3	0.222	0.217	0.217	0.214
1	0.220	0.220	0.215	0.211
4	-	-	0.218	0.223
5	-	-	0.212	0.218
10	-	-	-	0.224

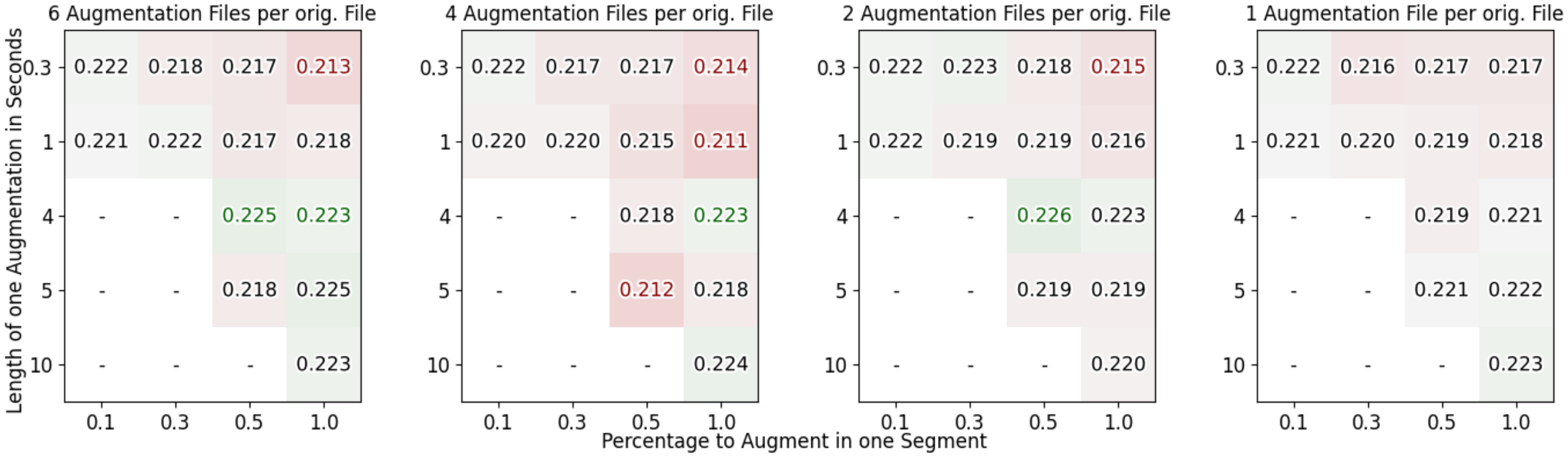
2 Augmentation Files per orig. File

	0.1	0.3	0.5	1.0
0.3	0.222	0.223	0.218	0.215
1	0.222	0.219	0.219	0.216
4	-	-	0.226	0.223
5	-	-	0.219	0.219
10	-	-	-	0.220

1 Augmentation File per orig. File

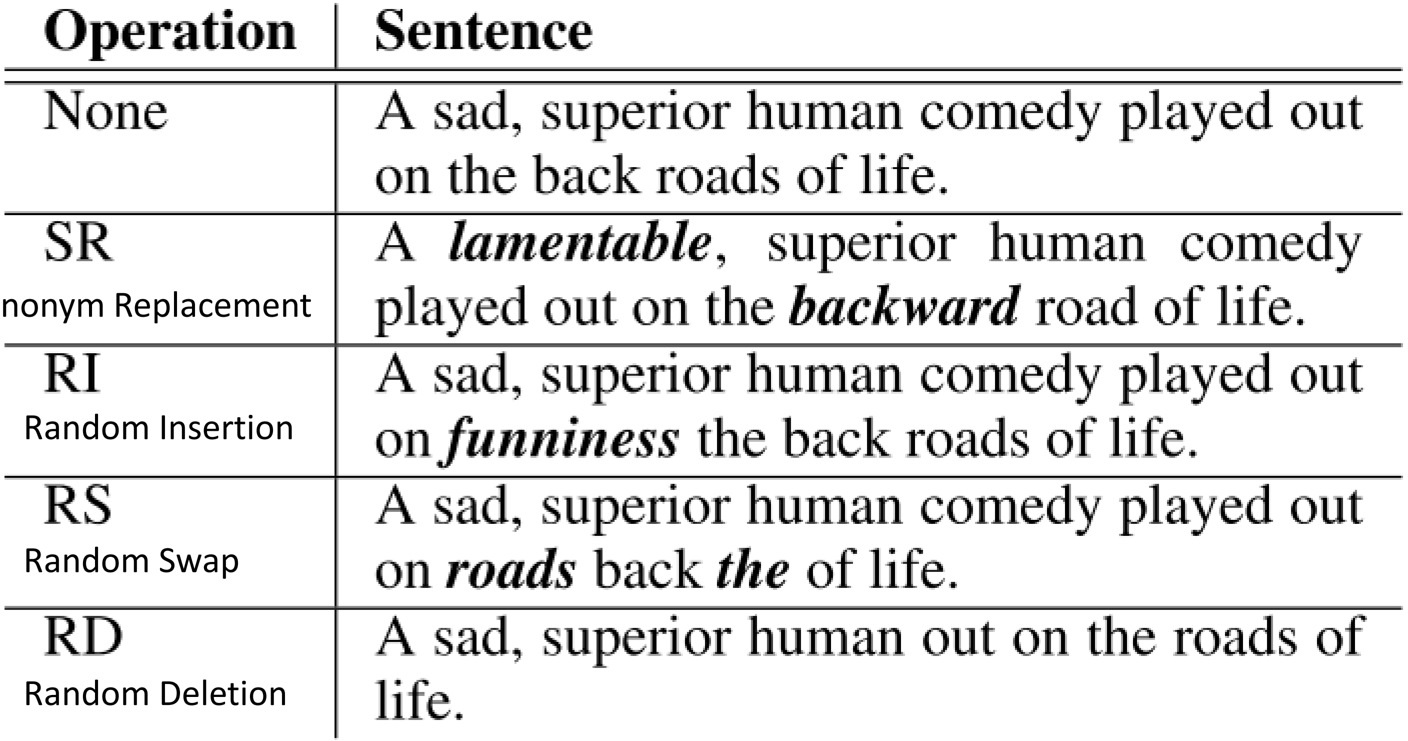
	0.1	0.3	0.5	1.0
0.3	0.222	0.216	0.217	0.217
1	0.221	0.220	0.219	0.218
4	-	-	0.219	0.221
5	-	-	0.221	0.222
10	-	-	-	0.223

Shifting Pitch



- Optimal: 2 Files per original File, 50% augmentation rate, length of 4 seconds each
- 0.5% enhancement compared to without augmentation

EDA



Operation	Sentence
None	A sad, superior human comedy played out on the back roads of life.
SR Synonym Replacement	A ***lamentable***, superior human comedy played out on the ***backward*** road of life.
RI Random Insertion	A sad, superior human comedy played out on ***funniness*** the back roads of life.
RS Random Swap	A sad, superior human comedy played out on ***roads*** back ***the*** of life.
RD Random Deletion	A sad, superior human out on the roads of life.

Shifting Pitch (SP)

Background Noise (BN)

Background Noise

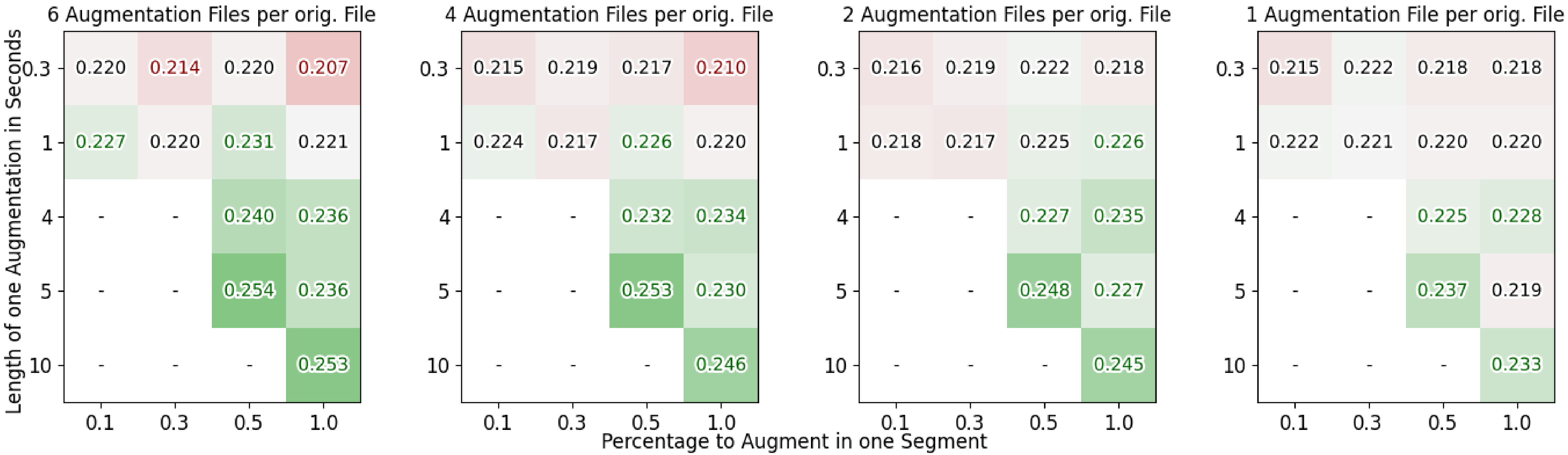
MUSAN

D. Snyder, G. Chen, and D. Povey, “Musan: A music, speech, and noise corpus,” arXiv preprint arXiv:1510.08484, 2015.

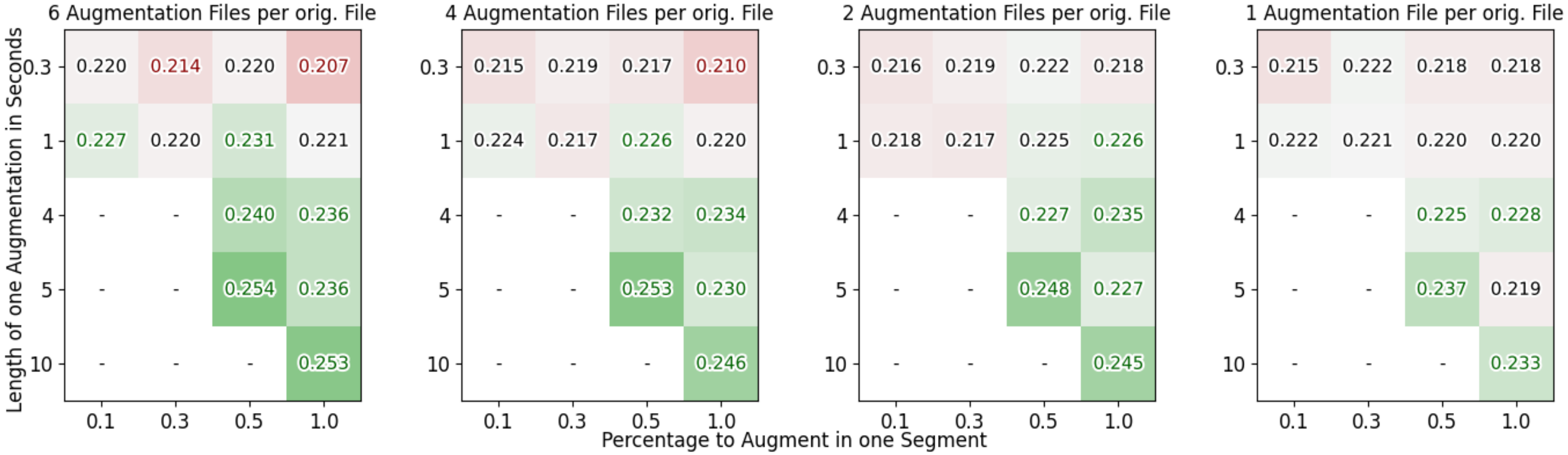
- 929 noise files
- Total duration ~6h
- Technical noises such as
  - Dialtones
  - Fax machine
- Ambient sounds such as
  - Car idling
  - Thunder/wind/rain
  - Paper rustling
  - Animal noises

- Random part of random noise file
- Scale noise that resulting SNRdB is in [0,30]

Background Noise



Background Noise



6 Augmentation Files per orig. File

	0.1	0.3	0.5	1.0
0.3	0.220	0.214	0.220	0.207
1	0.227	0.220	0.231	0.221
4	-	-	0.240	0.236
5	-	-	0.254	0.236
10	-	-	-	0.253

4 Augmentation Files per orig. File

	0.1	0.3	0.5	1.0
0.3	0.215	0.219	0.217	0.210
1	0.224	0.217	0.226	0.220
4	-	-	0.232	0.234
5	-	-	0.253	0.230
10	-	-	-	0.246

2 Augmentation Files per orig. File

	0.1	0.3	0.5	1.0
0.3	0.216	0.219	0.222	0.218
1	0.218	0.217	0.225	0.226
4	-	-	0.227	0.235
5	-	-	0.248	0.227
10	-	-	-	0.245

1 Augmentation File per orig. File

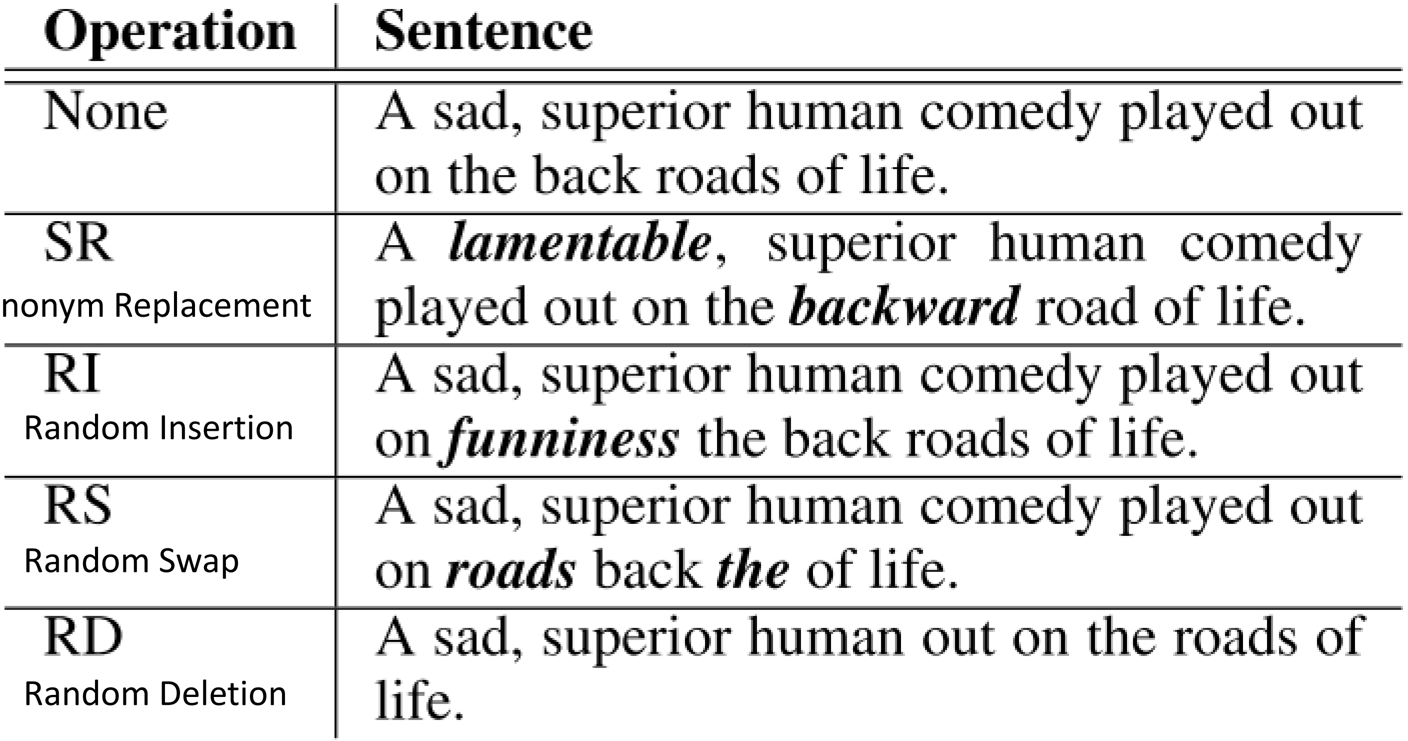
	0.1	0.3	0.5	1.0
0.3	0.215	0.222	0.218	0.218
1	0.222	0.221	0.220	0.220
4	-	-	0.225	0.228
5	-	-	0.237	0.219
10	-	-	-	0.233

- inserting background noise works best when only one noise sound is inserted
- Optimal: 6 Files per original File, 50% augmentation rate, length of 5 seconds each
- 3.3% enhancement compared to without augmentation
- worse with shorter augmentation length
  - Because of general shorter length or the chosen noise sounds?
  - Test again for 6 Files per original File, 100% augmentation rate, length of 0.3 seconds each
  - Only use files from MULAN with >=5seconds

    → significant better result, but still significant worse than without augmentation

    → important to use the right noise file

EDA



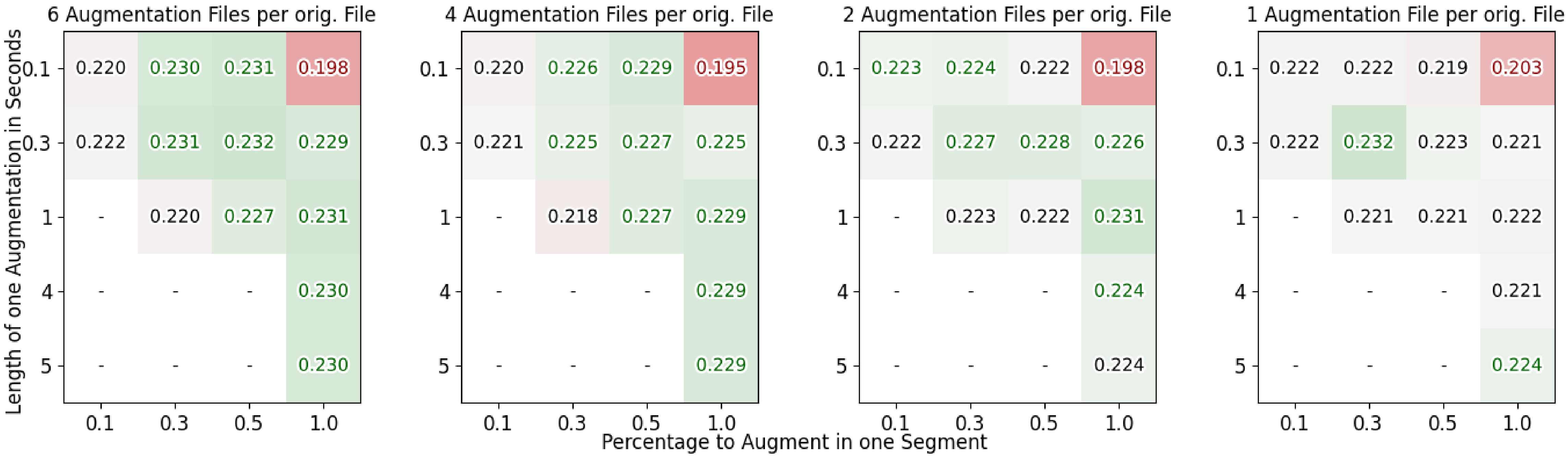
Operation	Sentence
None	A sad, superior human comedy played out on the back roads of life.
SR Synonym Replacement	A ***lamentable***, superior human comedy played out on the ***backward*** road of life.
RI Random Insertion	A sad, superior human comedy played out on ***funniness*** the back roads of life.
RS Random Swap	A sad, superior human comedy played out on ***roads*** back ***the*** of life.
RD Random Deletion	A sad, superior human out on the roads of life.

Shifting Pitch (SP)

Background Noise (BN)

Segment Swap (SeS)

Segment Swap



6 Augmentation Files per orig. File

	0.1	0.3	0.5	1.0
0.1	0.220	0.230	0.231	0.198
0.3	0.222	0.231	0.232	0.229
1	-	0.220	0.227	0.231
4	-	-	-	0.230
5	-	-	-	0.230

4 Augmentation Files per orig. File

	0.1	0.3	0.5	1.0
0.1	0.220	0.226	0.229	0.195
0.3	0.221	0.225	0.227	0.225
1	-	0.218	0.227	0.229
4	-	-	-	0.229
5	-	-	-	0.229

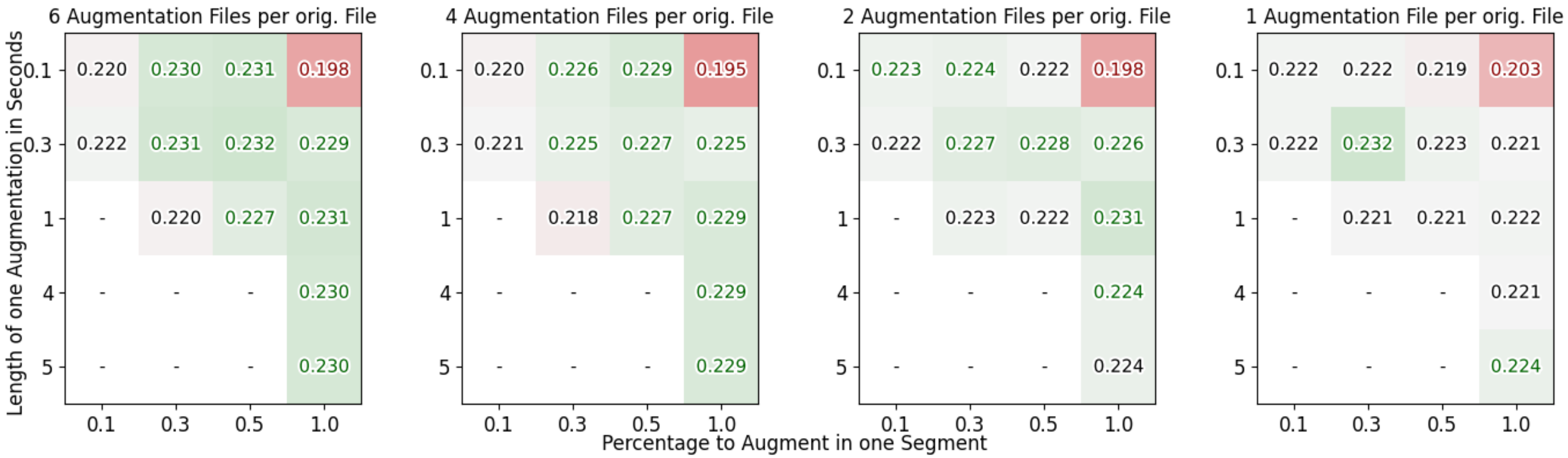
2 Augmentation Files per orig. File

	0.1	0.3	0.5	1.0
0.1	0.223	0.224	0.222	0.198
0.3	0.222	0.227	0.228	0.226
1	-	0.223	0.222	0.231
4	-	-	-	0.224
5	-	-	-	0.224

1 Augmentation File per orig. File

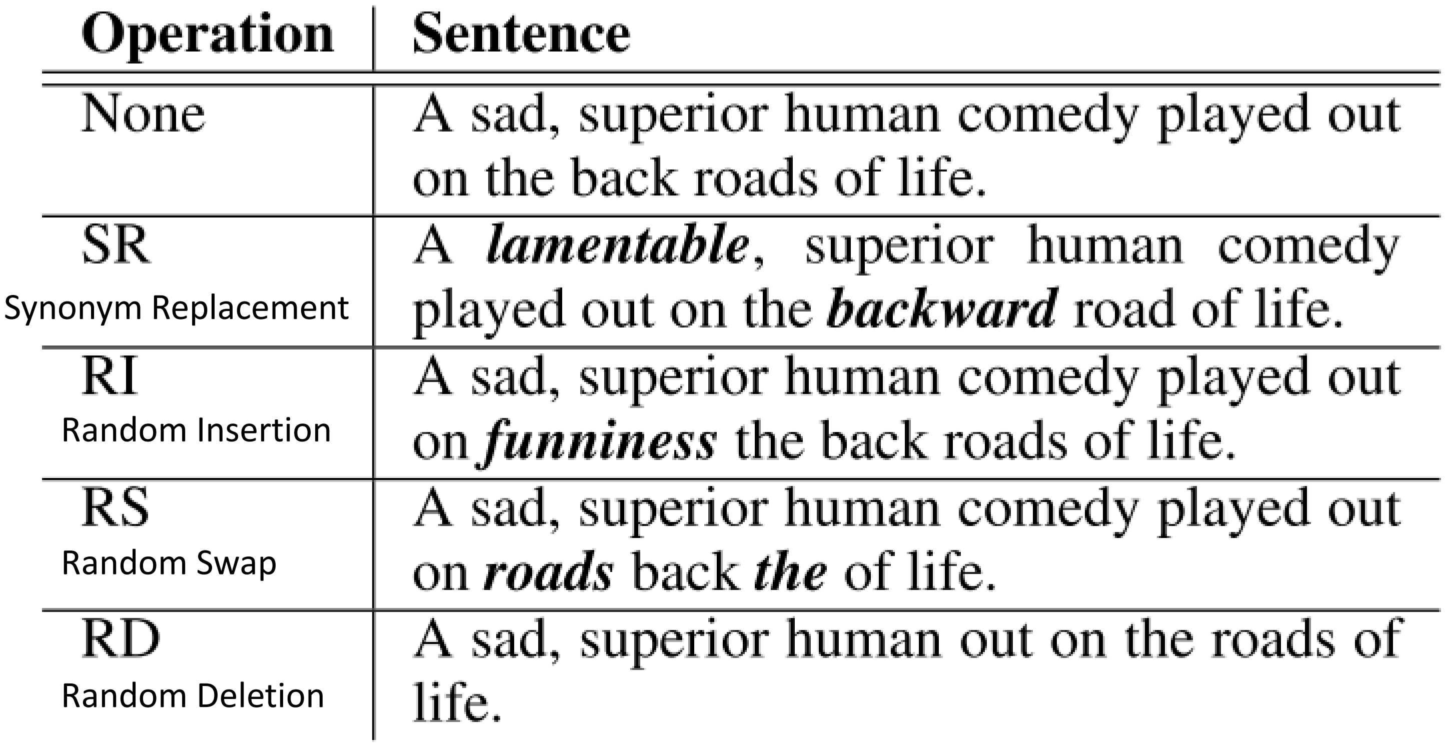
	0.1	0.3	0.5	1.0
0.1	0.222	0.222	0.219	0.203
0.3	0.222	0.232	0.223	0.221
1	-	0.221	0.221	0.222
4	-	-	-	0.221
5	-	-	-	0.224

Segment Swap

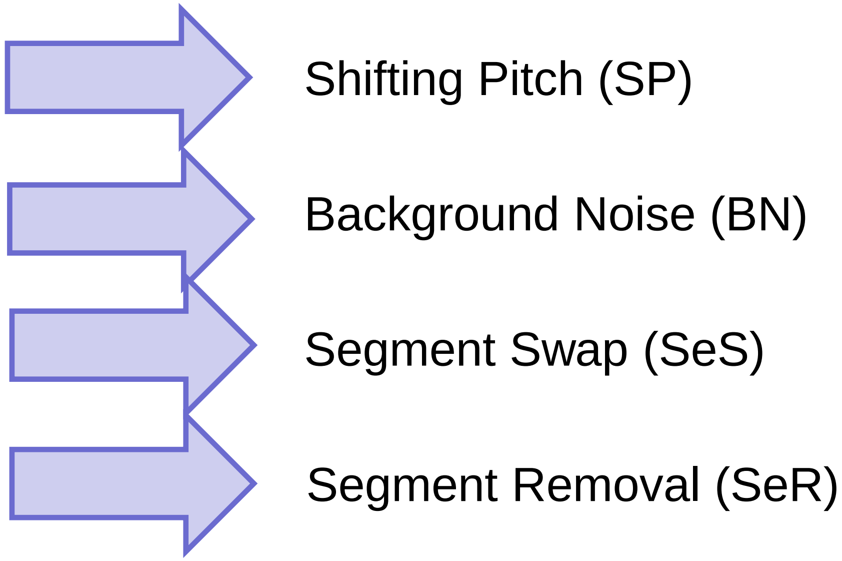


- one parameter combination with a significantly poorer outcome:
  100% augmentation and 0.1 seconds
  - insufficient duration of 0.1 seconds
  - related to the length of the vowels and consonants (length <0.3seconds)
- Optimal: 1 File per original File, 30% augmentation rate, length of 0.3 seconds each
  - using 6 files (with same Score) may not justify the increased computational overhead
- 1.1% enhancement compared to without augmentation

EDA



Operation	Sentence
None	A sad, superior human comedy played out on the back roads of life.
SR Synonym Replacement	A ***lamentable***, superior human comedy played out on the ***backward*** road of life.
RI Random Insertion	A sad, superior human comedy played out on ***funniness*** the back roads of life.
RS Random Swap	A sad, superior human comedy played out on ***roads*** back ***the*** of life.
RD Random Deletion	A sad, superior human out on the roads of life.



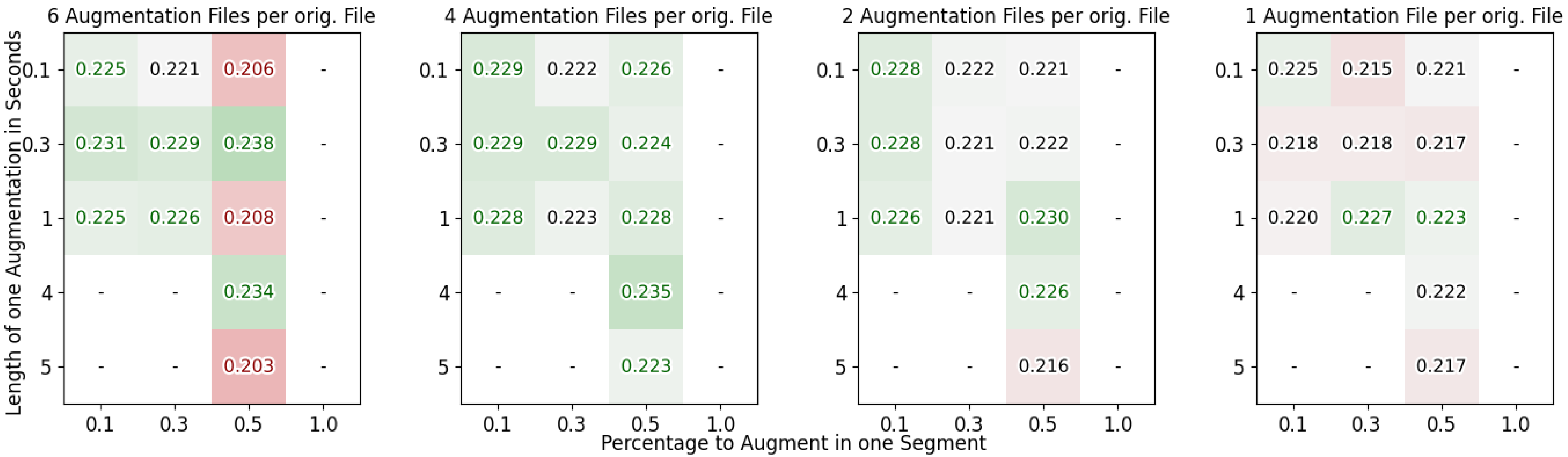
Shifting Pitch (SP)

Background Noise (BN)

Segment Swap (SeS)

Segment Removal (SeR)

Segment Removal



6 Augmentation Files per orig. File

	0.1	0.3	0.5	1.0
0.1	0.225	0.221	0.206	-
0.3	0.231	0.229	0.238	-
1	0.225	0.226	0.208	-
4	-	-	0.234	-
5	-	-	0.203	-

4 Augmentation Files per orig. File

	0.1	0.3	0.5	1.0
0.1	0.229	0.222	0.226	-
0.3	0.229	0.229	0.224	-
1	0.228	0.223	0.228	-
4	-	-	0.235	-
5	-	-	0.223	-

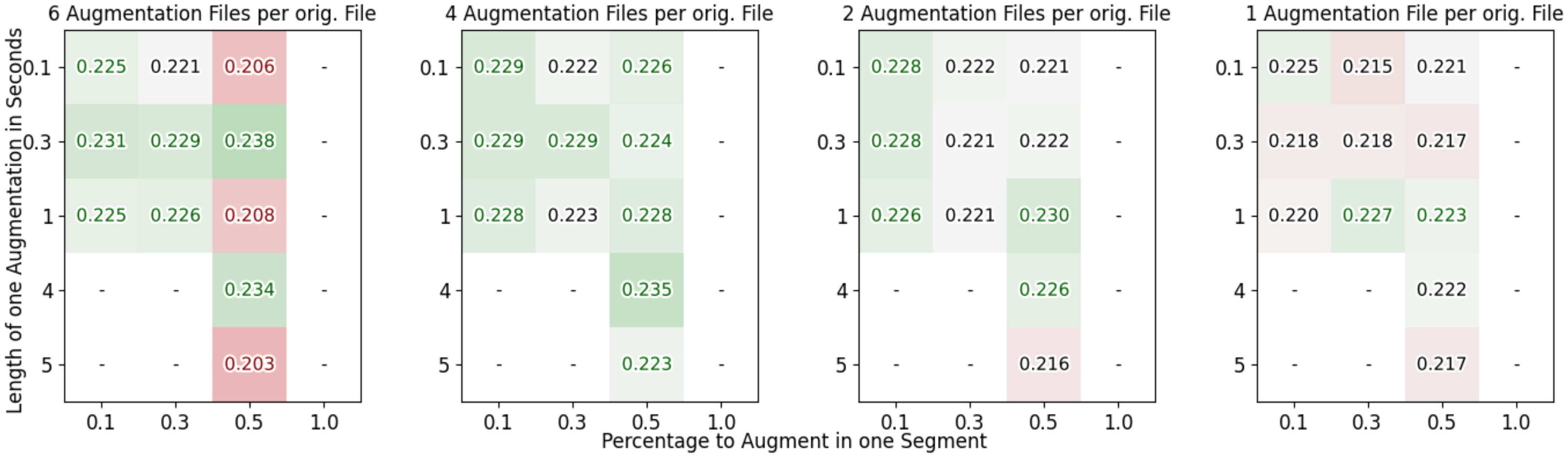
2 Augmentation Files per orig. File

	0.1	0.3	0.5	1.0
0.1	0.228	0.222	0.221	-
0.3	0.228	0.221	0.222	-
1	0.226	0.221	0.230	-
4	-	-	0.226	-
5	-	-	0.216	-

1 Augmentation File per orig. File

	0.1	0.3	0.5	1.0
0.1	0.225	0.215	0.221	-
0.3	0.218	0.218	0.217	-
1	0.220	0.227	0.223	-
4	-	-	0.222	-
5	-	-	0.217	-

Segment Removal



6 Augmentation Files per orig. File

	0.1	0.3	0.5	1.0
0.1	0.225	0.221	0.206	-
0.3	0.231	0.229	0.238	-
1	0.225	0.226	0.208	-
4	-	-	0.234	-
5	-	-	0.203	-

4 Augmentation Files per orig. File

	0.1	0.3	0.5	1.0
0.1	0.229	0.222	0.226	-
0.3	0.229	0.229	0.224	-
1	0.228	0.223	0.228	-
4	-	-	0.235	-
5	-	-	0.223	-

2 Augmentation Files per orig. File

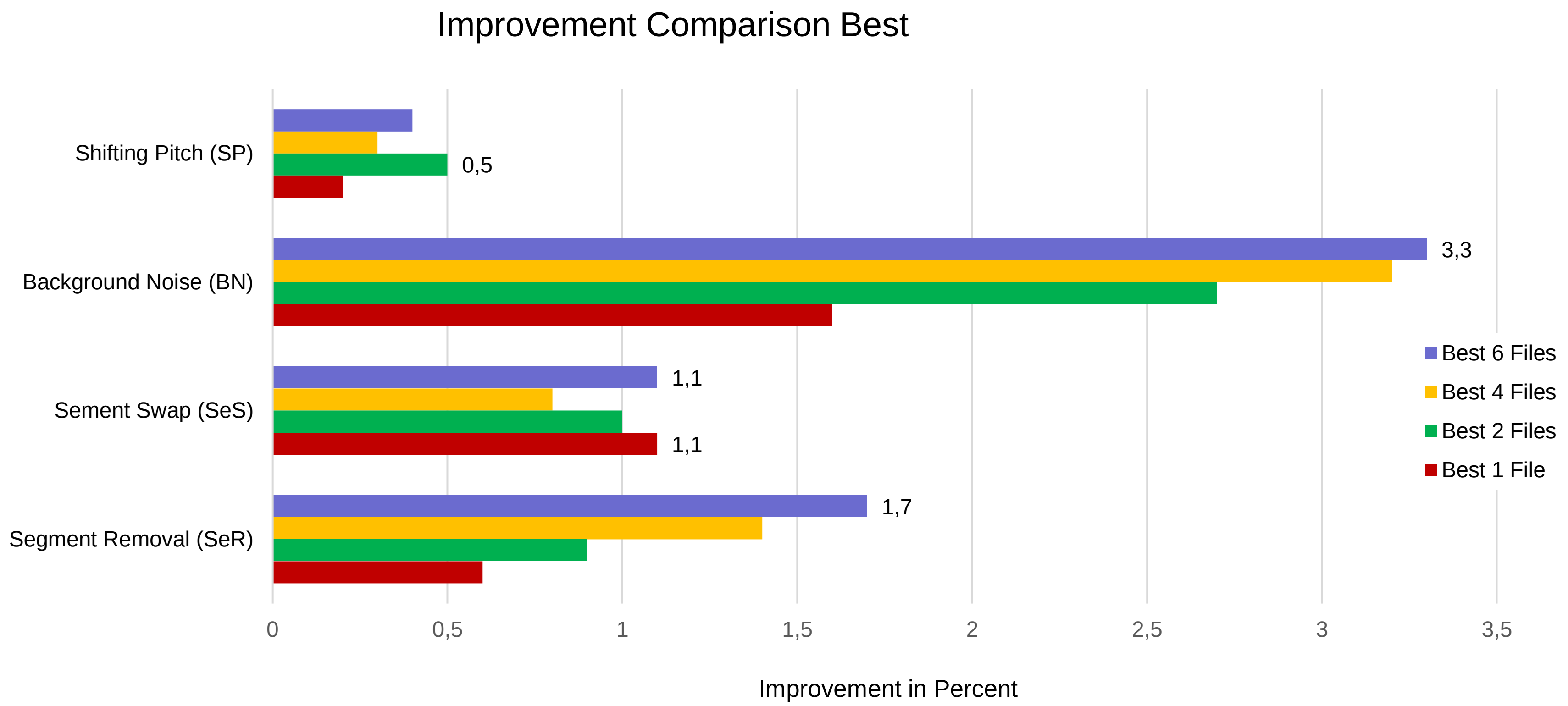
	0.1	0.3	0.5	1.0
0.1	0.228	0.222	0.221	-
0.3	0.228	0.221	0.222	-
1	0.226	0.221	0.230	-
4	-	-	0.226	-
5	-	-	0.216	-

1 Augmentation File per orig. File

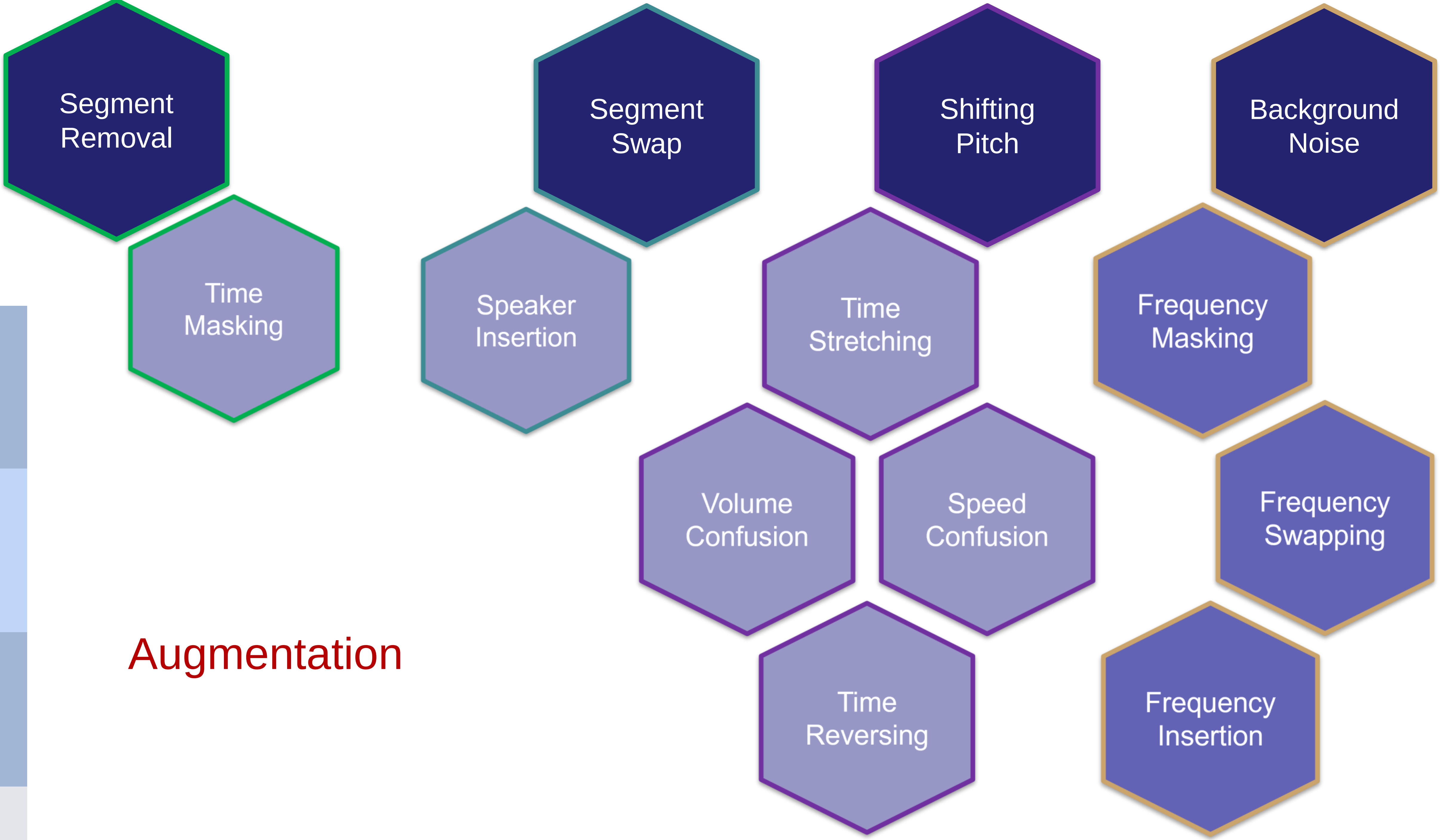
	0.1	0.3	0.5	1.0
0.1	0.225	0.215	0.221	-
0.3	0.218	0.218	0.217	-
1	0.220	0.227	0.223	-
4	-	-	0.222	-
5	-	-	0.217	-

- Optimal: 6 Files per original File, 50% augmentation rate, length of 0.3 seconds each
- 1.7% enhancement compared to without augmentation
- For 6 Files per original File, 50% augmentation rate there are three significant worse results
  - Only results with significant worse performance
  - Conversely not for 4, 2 or 1 Files per original File

Results Main Methods



Augmentation



Segment Removal

Time Masking

Segment Swap

Speaker Insertion

Shifting Pitch

Time Stretching

Volume Confusion

Speed Confusion

Time Reversing

Background Noise

Frequency Masking

Frequency Swapping

Frequency Insertion

Time Masking

- Similar to Segment Removal
- but the interval is not removed
- instead, it is replaced by zeros

- Used values for Hyperparameters:
  - $n_{augFiles} = 6$
  - $\propto = 0.5$
  - $l_{aug} = 0.3$

Speaker Insertion

- The specific interval is replaced by another random interval from a different speaker
- Speaker is of the same class (hence, the same dialect).

- Used values for Hyperparameters:
  - $n_{augFiles} = 1$
  - $\propto = 0.3$
  - $l_{aug} = 0.3$

Time Stretching & Speed Confusion

- Time Stretching
  - Intervals are time stretched within the range of [0.8, 1.2]
  - Pitch remains unchanged

- Speed Confusion
  - Similar to TS
  - But Pitch changes too
  - To archive that, the Interval gets resampled, but saved with the original sample rate
  - newSamplingRate=rate*oldSamplingRate,rate∈[0.8,1.2]

Volume Confusion & Time Reversing

- Volume Confusion
  - Peak of the segment is set to a value within the range [0.2, 0.8]

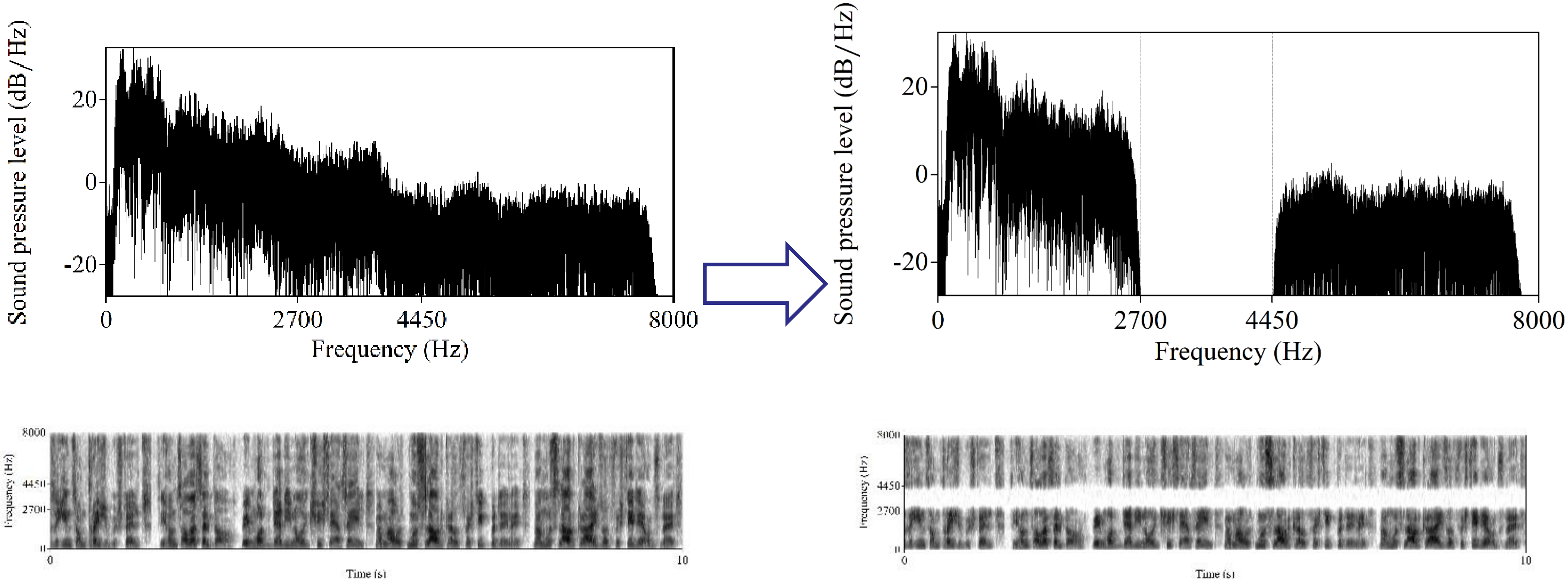
- Time Reversing
  - Order of the samples in the Interval gets reversed

- Used values for Hyperparameters:
  - $n_{augFiles} = 2$
  - $\propto = 0.5$
  - $l_{aug} = 4.0$

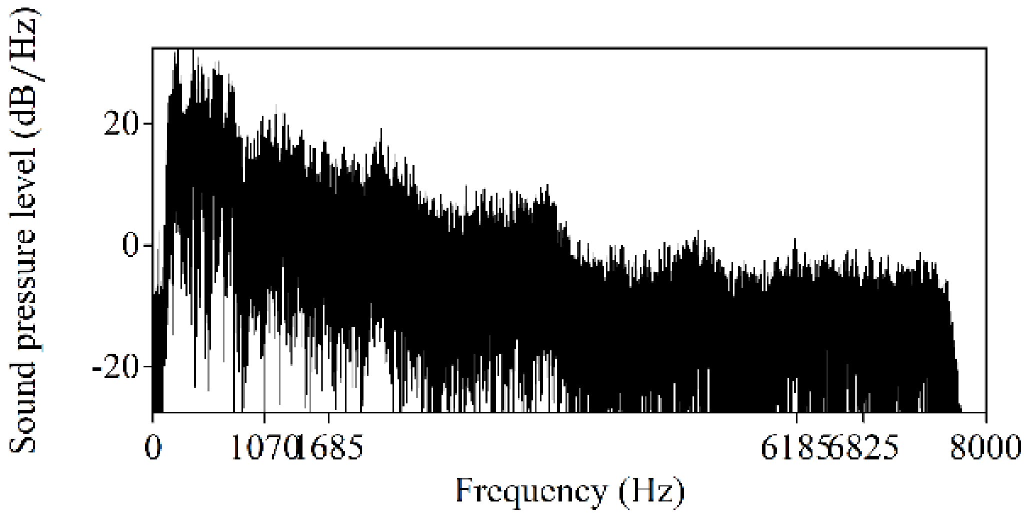
Frequency manipulation

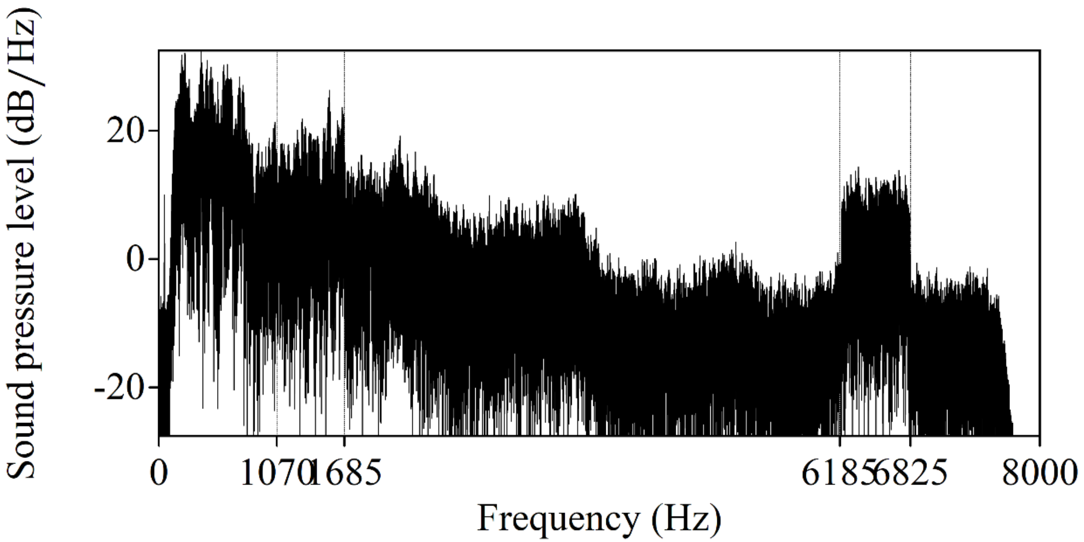
- Used values for Hyperparameters:
  - $n_{augFiles} = 6$
  - $\propto$ = not needed
  - l_{aug} = not needed

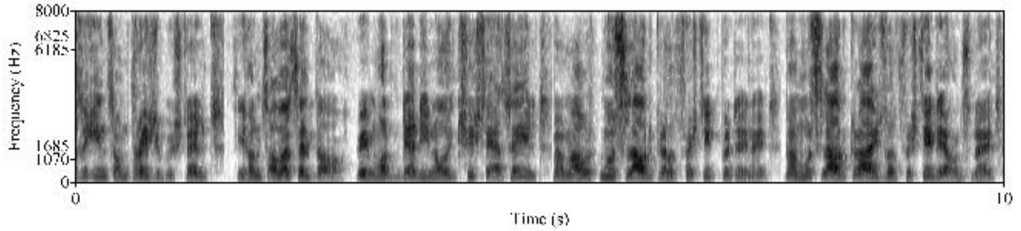
Frequency Masking

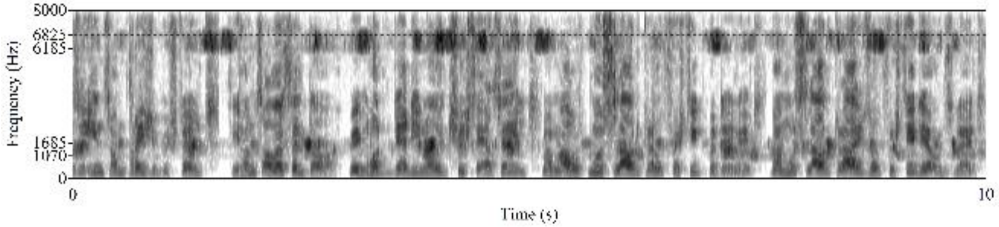


Frequency Insertion

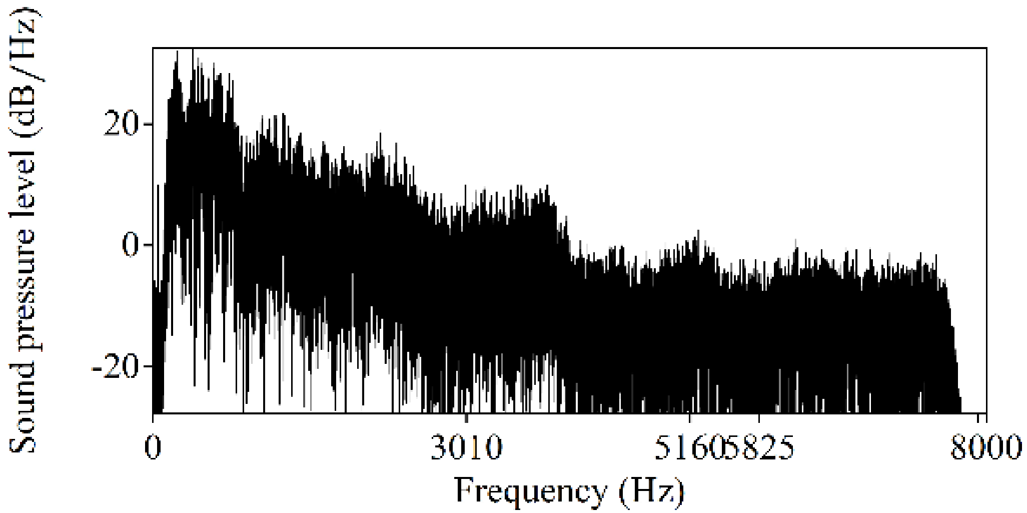


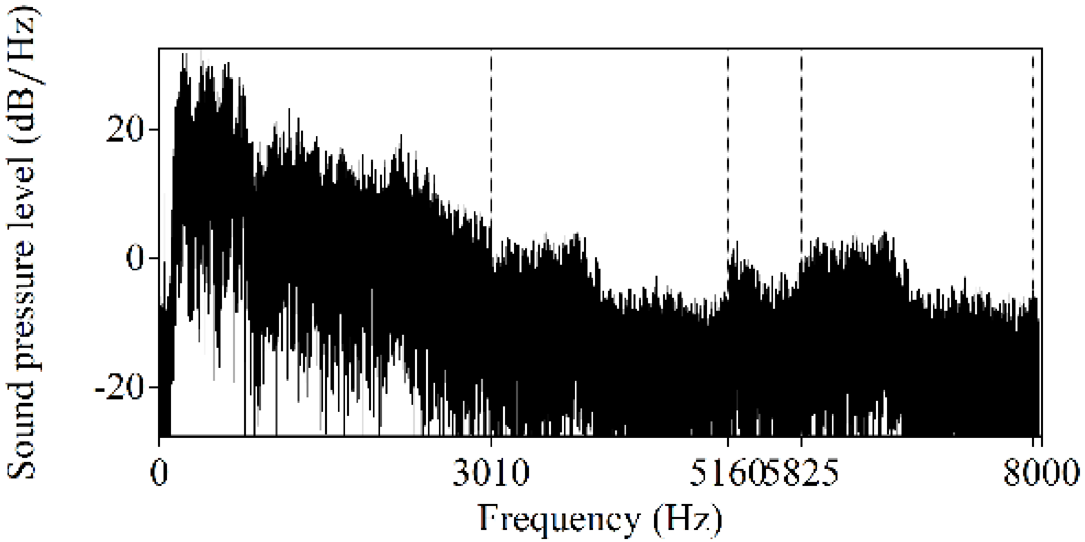


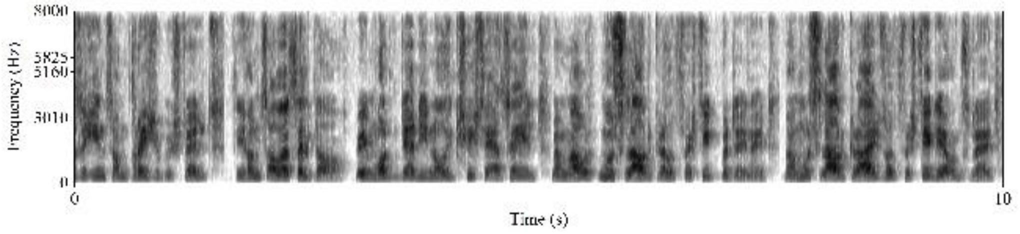


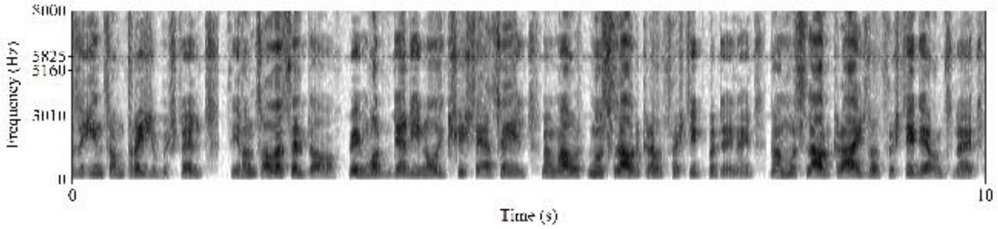


Frequency Swapping

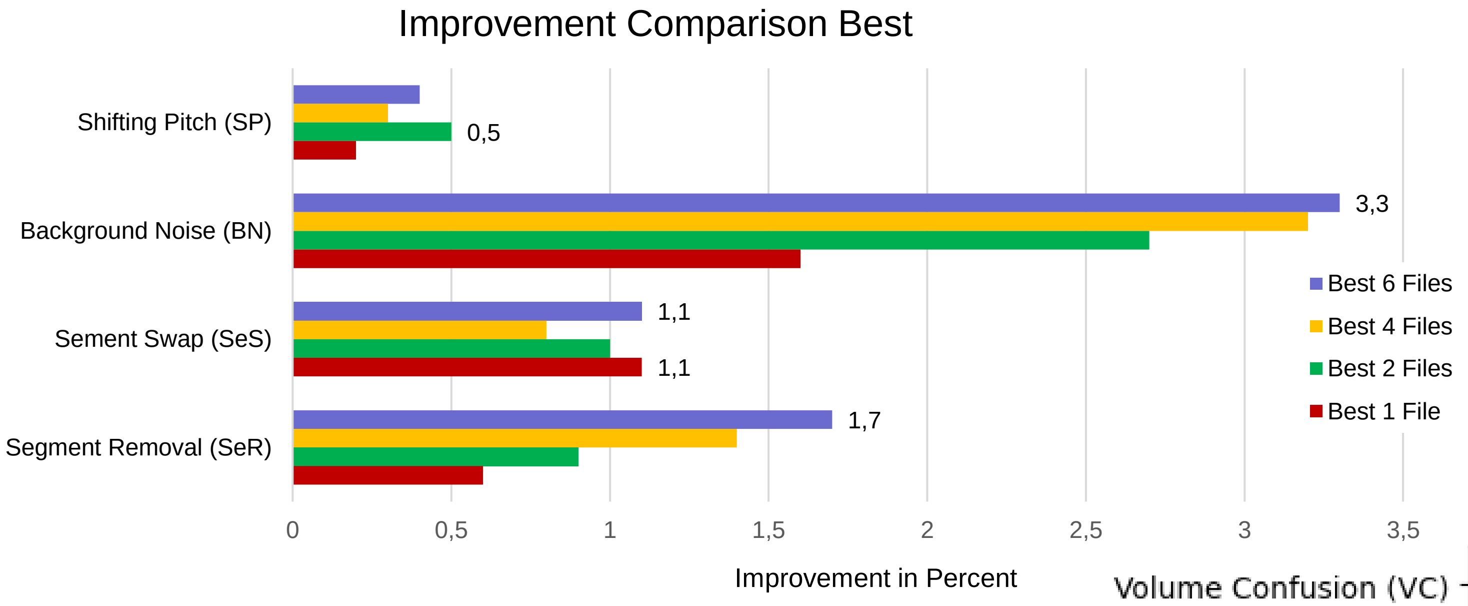


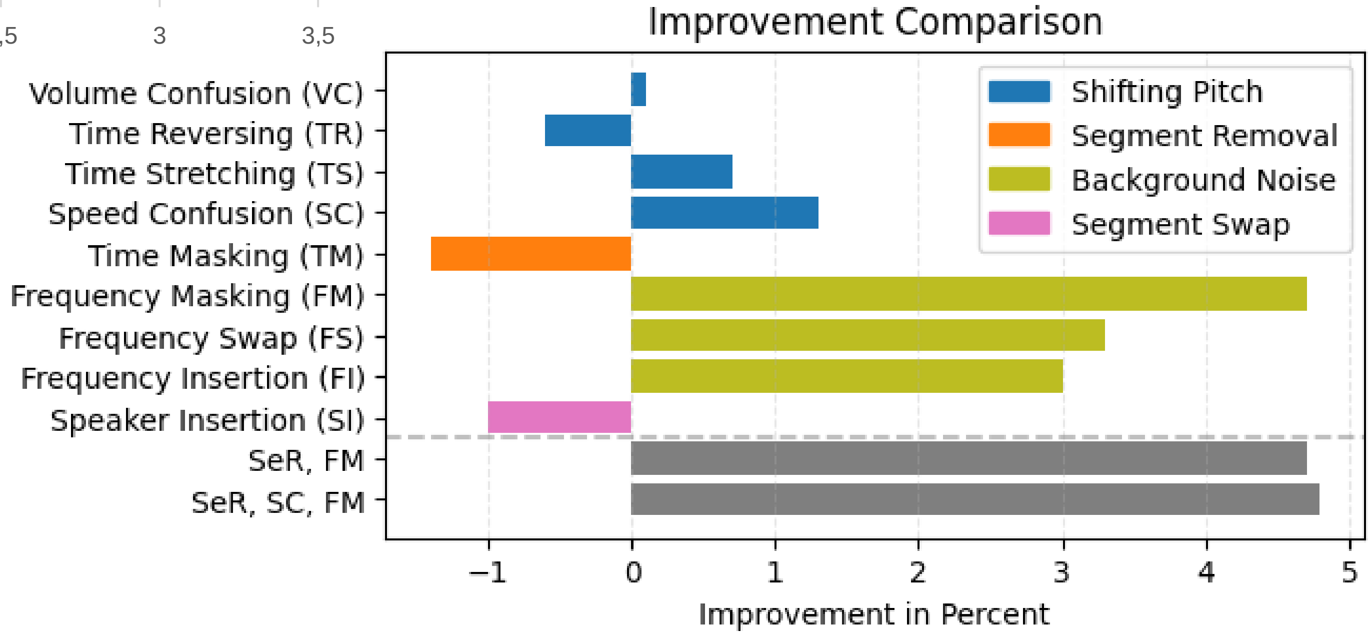






Results of all Methods





Conlusion

- Best Method is Frequency Masking
- 4.7% better than without augmentation (from 0.221 to 0.268)
- Generally, all methods that are masking frequencies yield the best results
- Can add Segment Removal without performance loss to reduce computation effort

Thanks for your attention!
Any Questions?