TEAM IMPORTED

Michele Giambelli

Korhan Erdoğdu

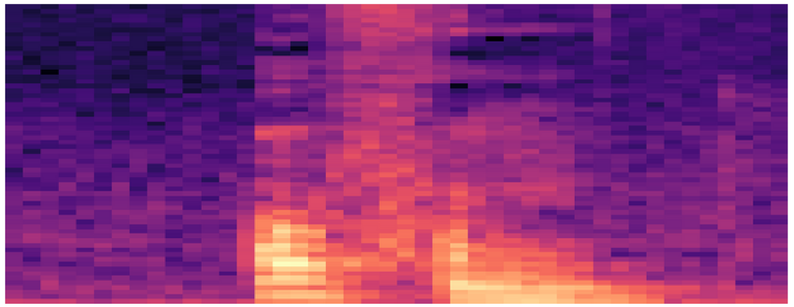
Mahmut Öztürk

Barış Bakırdöven

OUR SYSTEM

How to detect the complete speech commands

- **Windowing Approach:** Created 44-frame windows (1.1 seconds) across longer recordings, sliding the window with a 3-frame step size.
- **Keyword Detection**: Applied the CNN classifier to each windowed segment for keyword detection.
- **Obtain the complete speech command:** Going backward in the predicitons and combine each “command” keyword with the before “device” keyword.



3_speech_true_Radio_an\844-887.png	other
3_speech_true_Radio_an\847-890.png	Radio
3_speech_true_Radio_an\850-893.png	Radio
3_speech_true_Radio_an\853-896.png	Radio
3_speech_true_Radio_an\856-899.png	Radio
3_speech_true_Radio_an\859-902.png	Radio
3_speech_true_Radio_an\862-905.png	Radio
3_speech_true_Radio_an\865-908.png	Radio
3_speech_true_Radio_an\868-911.png	Radio
3_speech_true_Radio_an\871-914.png	other
3_speech_true_Radio_an\874-917.png	other
3_speech_true_Radio_an\877-920.png	other
3_speech_true_Radio_an\880-923.png	other
3_speech_true_Radio_an\883-926.png	other
3_speech_true_Radio_an\886-929.png	other
3_speech_true_Radio_an\889-932.png	other
3_speech_true_Radio_an\892-935.png	an
3_speech_true_Radio_an\895-938.png	an
3_speech_true_Radio_an\898-941.png	an
3_speech_true_Radio_an\901-944.png	other

filename	command	timestamp
3_speech_true_Radio_an	Radio an	22.25

MINIMIZE THE LOSS FUNCTION

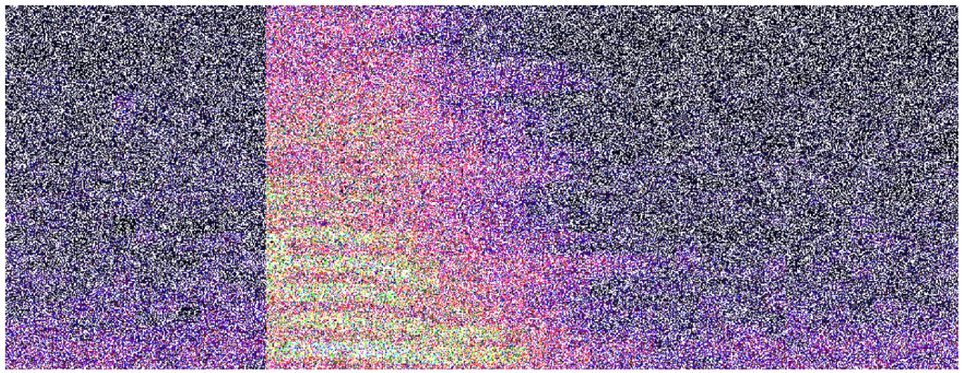
- Post-Processing:
  - **Smoothing Predictions:** To each prediction applied neighborhood voting to smooth prediction sequences.
  - **Filtering:** Compare the length of the segments with a threshold(2) and if it is shorter than 2 change the labels to “Other.

1163_speech_true_Alarm_aus\337-380.png	Fernseher	Fernseher
1163_speech_true_Alarm_aus\340-383.png	Fernseher	Fernseher
1163_speech_true_Alarm_aus\343-386.png	other	Fernseher
1163_speech_true_Alarm_aus\346-389.png	Fernseher	Fernseher
1163_speech_true_Alarm_aus\349-392.png	Fernseher	Fernseher
1163_speech_true_Alarm_aus\352-395.png	Fernseher	Fernseher
1163_speech_true_Alarm_aus\355-398.png	other	other
1163_speech_true_Alarm_aus\358-401.png	other	other

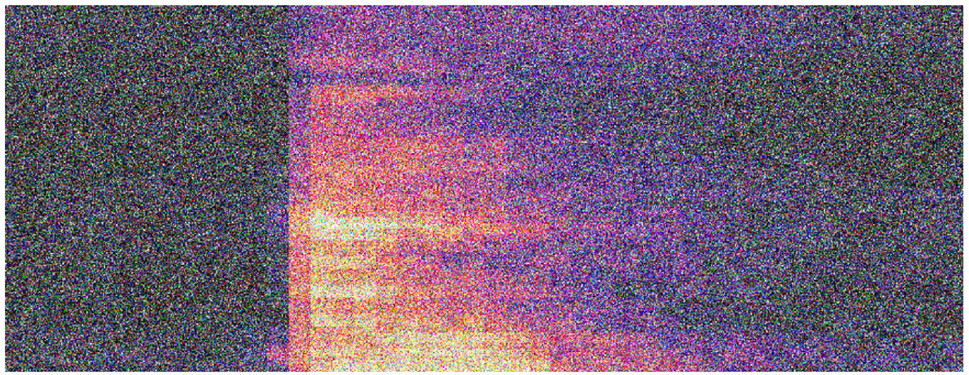
00c6e9ce33\373-416.png	other	other	other
00c6e9ce33\376-419.png	Alarm	Alarm	Alarm
00c6e9ce33\379-422.png	Alarm	Alarm	Alarm
00c6e9ce33\382-425.png	Alarm	Alarm	Alarm
00c6e9ce33\385-428.png	Alarm	Alarm	Alarm
00c6e9ce33\388-431.png	Alarm	Alarm	Alarm
00c6e9ce33\391-434.png	Alarm	Alarm	Alarm
00c6e9ce33\394-437.png	Alarm	Alarm	Alarm
00c6e9ce33\397-440.png	other	other	other
00c6e9ce33\400-443.png	other	other	other
00c6e9ce33\403-446.png	other	other	other
00c6e9ce33\406-449.png	other	other	other
00c6e9ce33\409-452.png	other	other	other
00c6e9ce33\412-455.png	other	other	other
00c6e9ce33\415-458.png	aus	aus	aus
00c6e9ce33\418-461.png	aus	aus	aus
00c6e9ce33\421-464.png	Staubsauger	Staubsauger	other
00c6e9ce33\424-467.png	other	other	other

MINIMIZE THE LOSS FUNCTION

- Data Augmentation:
  - Added noise and dropped pixels from mel spectrograms to enhance sensitivity to keywords.



Dropout



Noisy

- Hyperparameter Tuning:
  - Experimented with various parameters to minimize the custom cost function.

RESULTS

- Accuracy:

Naïve Baseline	CNN	CNN hyperparameter tuning	CNN + Data agumentation
9%	94.6%	90%	87.02%

- Loss function:

Naïve Baseline	CNN
430	− 221