Algorithms and Evaluation Guide v1.1

Femtosense includes the following pre-trained algorithms as part of its evaluation kit:

- Single-microphone “Hey Snips” Wake Word Detection
- Single-microphone AI Noise Reduction

Model Name	Description
WWDSNIPS108161	Wake Word Detection “Hey Snips” 1 microphone 8 kHz sampling rate 16 ms hop-size Version 1.0
WWDSNIPS116161	Wake Word Detection “Hey Snips” 1 microphone 16 kHz sampling rate 16 ms hop-size Version 1.0
AINRGP11608161	AI Noise Reduction “General Purpose” 1 microphone 16 kHz sampling rate 8 ms hop-size 16 ms algorithmic latency Version 1.0
AINRGP11604081	AI Noise Reduction “General Purpose” 1 microphone 16 kHz sampling rate 4 ms hop-size 8 ms algorithmic latency Version 1.0

Table of Contents

1. Algorithms

1.1 Wake Word Detection Algorithms

These wakeword detection algorithm is trained to recognize the phrase “Hey Snips.” The model input is a sequence of raw waveform frames, and its output is a sequence of probabilities of the presence of the keyword at each frame. The model has been trained against indoor environmental noise, competing speech, and room reverberance. The model audio input uses an INT16 PCM format.

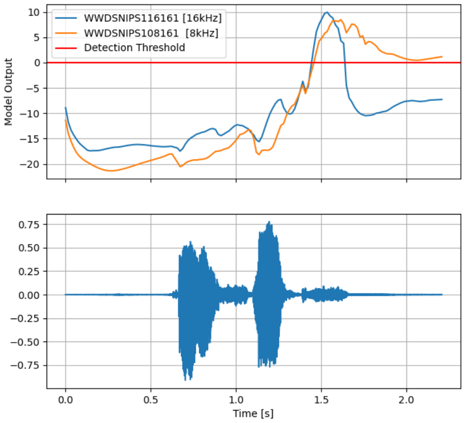
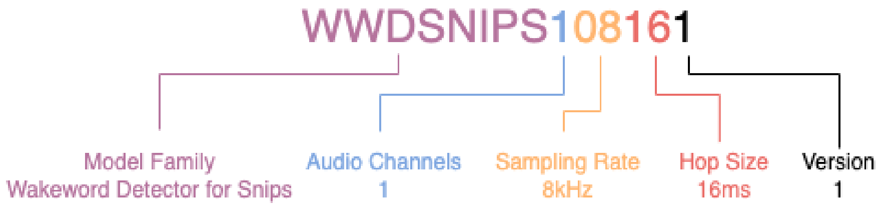


Figure 1: Wakeword detection algorithms return scalar values for each input frame of audio (every 16ms). When the output exceeds the detection threshold (default 0), it is interpreted as a positive detection of the keyword.

Variants exist for 16kHz and 8kHz sampling rates.

Model Naming Convention Example

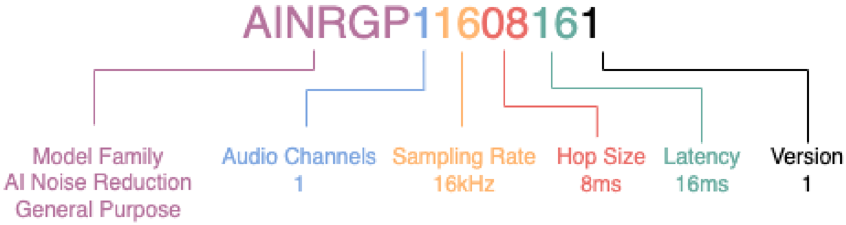


- Audio Channels: 1
- Input audio sampling rate: 8kHz
- Hop size: 16ms
- Algorithm version: 1.0

1.2 AI Noise Reduction Algorithms

The general purpose AI Noise Reduction algorithm removes background noise while preserving human speech. Speech can be in any language. The model's input is a sequence of noisy raw waveform frames, and its output is a sequence of enhanced waveform frames. The model input and output audio uses an INT16 PCM format.

Model Naming Convention Example



- Audio Channels: 1
- Input/output audio sampling rate: 16kHz
- Hop size: 8ms
- Algorithmic latency: 16ms
- Algorithm version: 1.0

2. Proposed Evaluation

2.1 Wake Word Detection

Our algorithm detects `Hey Snips` in a large variety of environments. We recommend evaluating the algorithm in the following conditions:

- Noise Environments: music noise, competing speech
- Signal to Noise Ratios: 3dB SNR or higher
- Distance: Play source audio at a distance from 0 to 2 meters from the microphone.

The model was trained on a small dataset of speakers with American accents. Performance may degrade for speakers with other accents. To aid in the evaluation, we provide a small set of validation audio files for testing purposes. These audio samples were not used during the training of the model.

Specifications:

- Audio In:
    - 8 kHz Sampling Rate
    - Monaural
    - 16 bits (pcm)
- Output: scalar probability of keyword

Testing should be conducted in a non-reverberant environment when mixing with noise, otherwise the effective SNRs levels will be lower.

2.2 AI Noise Reduction

Our algorithm removes background noise while preserving the speech. We recommend evaluating the algorithm in the following conditions:

- Noise Environments: car noise, babble noise (restaurant/coffee background), transient sounds
- Signal to Noise Ratios: The algorithm should provide good performance above -3 dB SNR. The algorithm should work without voice distortions above 0 dB SNR.
- Distance: Play source audio at a distance from 0 to 3 meters from the microphone.

We suggest experiencing the algorithm while wearing an Active Noise Cancellation (ANC) headset to reduce the noise and speech that reaches the user's ears directly from the speaker. This approach replicates the usage scenario where our algorithm would be combined with ANC in earbuds to mitigate the direct path. For assistive hearing devices, patients with hearing loss would experience less of a direct path than people with normal hearing.

Specifications:

- Audio In:
  - 16 kHz Sampling Rate
  - Monaural
  - 16 bits (PCM)
- Audio out:
  - 16 kHz Sampling Rate
  - Monaural
  - 16 bits (PCM)

Our algorithm is not trained for a specific microphone model. For reference, the microphone we used for testing had the following specifications.

- MEMS microphone
- SNR: 67.5 dB
- AOP: 123 dB
- Sensitivity: -38 ± 3 dBFS

Testing should be conducted in a non-reverberant environment, otherwise the effective SNRs levels will be lower.

Model Performance:

The model works well down to 0dB SNR without voice distortion. At -3dB, the voice is still preserved but the user can expect mild distortion.

We measured the performance of our algorithm with six different metrics across SNR levels (-6dB, -3dB, 0dB, +3dB, +6dB) for car and babble speech noise environments. We use both intrusive metrics, and perceptual metrics.

Intrusive metrics use both the clean target speech file and the enhanced audio. We use the following intrusive metrics,

- SISDR as a measure of the amount of noise removed by the algorithm
- PESQ as a measure of the speech quality of the processed audio
- STOI as a measure of the speech intelligibility improvement by the algorithm

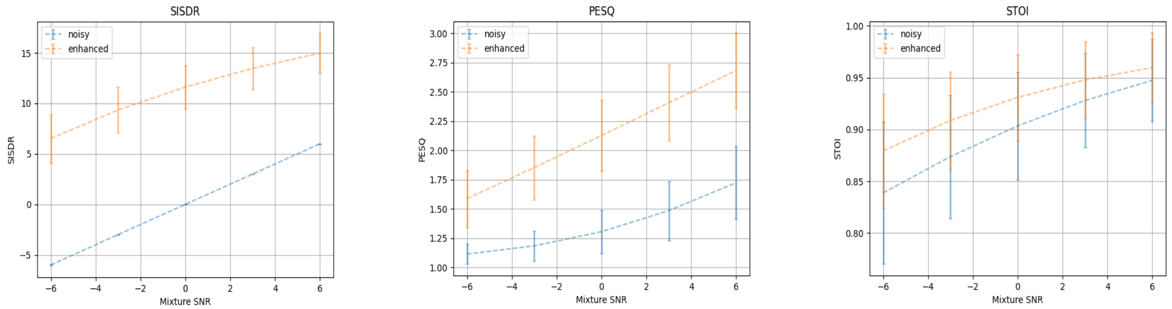


Figure 2: Improvements in intrusive metrics from AIMRGP11608161 with a wide variety of car noises from the Vehicle Interior Sounds Dataset across SNR levels

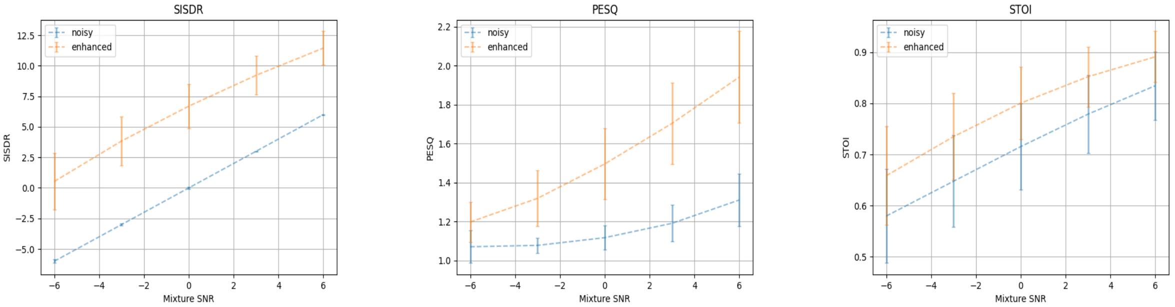


Figure 3: Improvements in intrusive metrics from AIMRGP11608161 with a wide variety of speech babble noises from the WHAM! Dataset across SNR levels

Perceptual metrics to model the subjective human experience of quality. They are generated by the models described in the DNSMOS P.835 paper by Microsoft and are reportedly correlated highly with human Mean Opinion Scores. For a detailed explanation about the scale of these metrics, please refer to this paper. Higher scores means higher audio quality. We use the following perceptual metrics,

- OVR measures the overall audio quality
- BAK measures the amount of noise removal
- SIG measures the speech quality

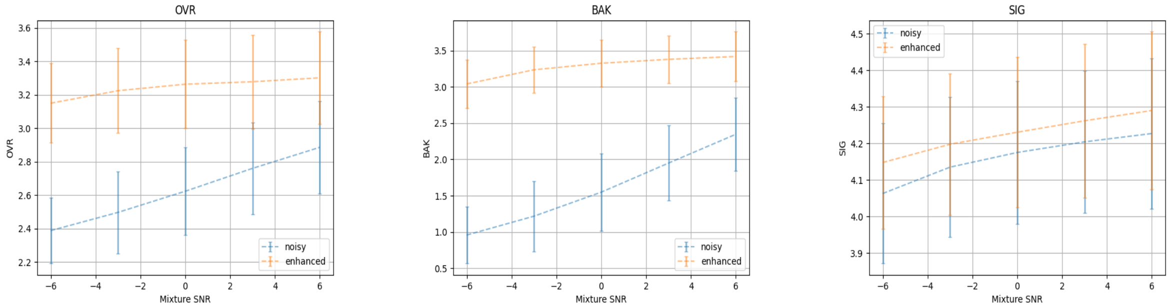


Figure 4: Improvements in perceptual metrics from AIMRGP11608161 with a wide variety of car noises from the Vehicle Interior Sounds Dataset across SNR levels

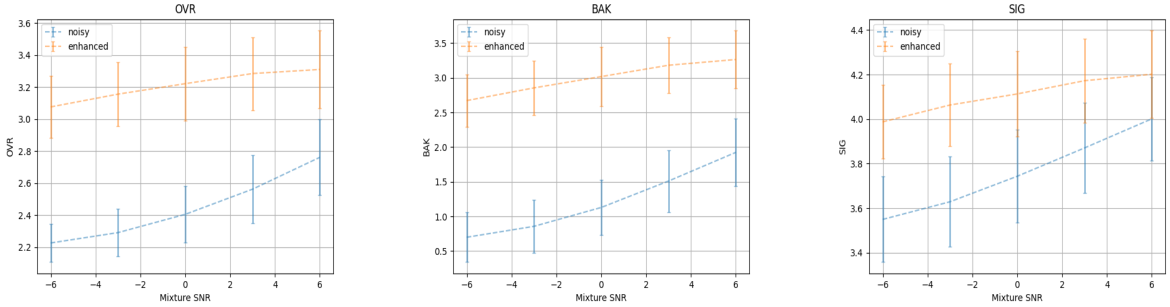


Figure 5: Improvements in perceptual metrics from AIMRGP11608161 with a wide variety of speech babble files from the WHAM! Dataset across SNR levels

3. Change Log

Version	Release Date	Description
1.0	2023-04-09	Initial release
1.1	2023-05-03	Add model performance graphs and reports