An Efferent-inspired Auditory Model Front-end for Speech Recognition

Chia-ying Lee, James Glass and Oded Ghitza*

MIT Computer Science and Artificial Intelligence Lab, Cambridge, MA, USA
*Boston University Hearing Research Lab, Boston, MA, USA

Motivation

- **Human v.s. Automatic Speech Recognizers (ASRs)**
  - Humans are particularly good at dealing with previously unseen noise or dynamic noises.

Motivation

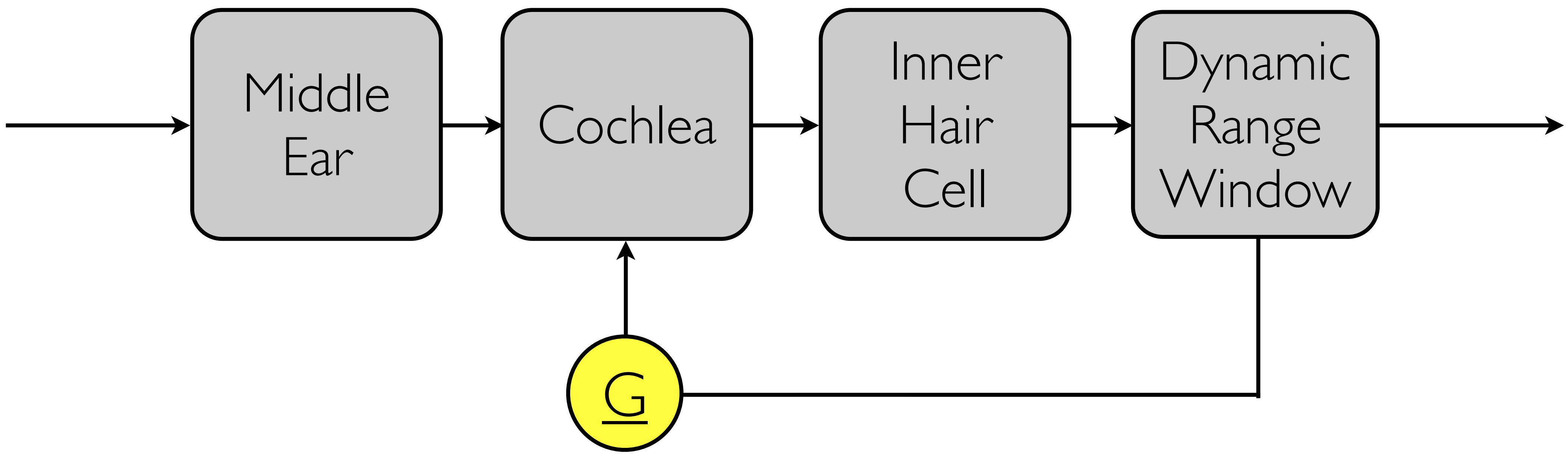
- **Human v.s. Automatic Speech Recognizers (ASRs)**
  - Humans are particularly good at dealing with previously unseen noise or dynamic noises.
- **Mounting evidence of the role of efferent-feedback in mammalian auditory systems**
  - Operating point of the cochlea is regulated by background noise
  - Results in stable internal representations

Motivation

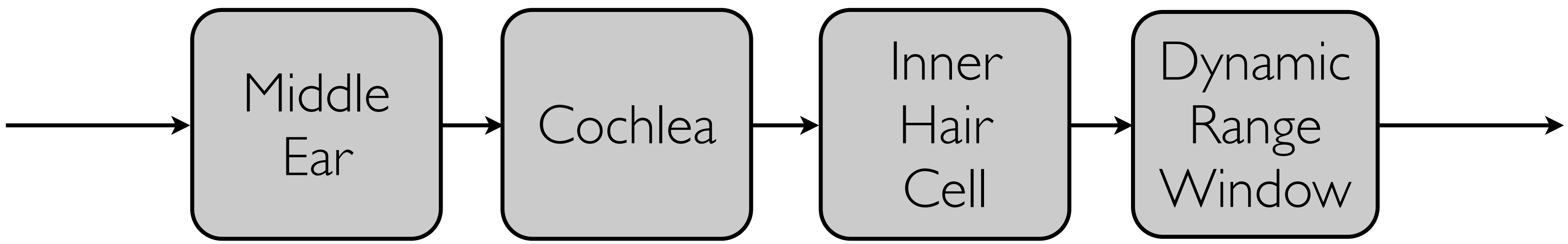
- **Human v.s. Automatic Speech Recognizers (ASRs)**
  - Humans are particularly good at dealing with previously unseen noise or dynamic noises.
- **Mounting evidence of the role of efferent-feedback in mammalian auditory systems**
  - Operating point of the cochlea is regulated by background noise
  - Results in stable internal representations
- **Explore potential use of a feedback mechanism for ASR**
  - Use a MOC efferent-inspired auditory model as an ASR front-end

An Efferent-inspired Auditory Model

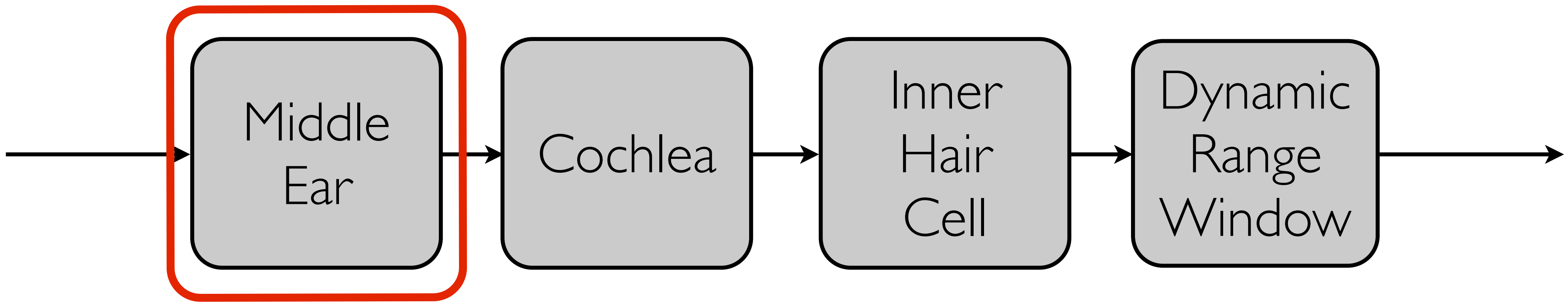
- Messing et al., 2009



Model of Ascending Pathway

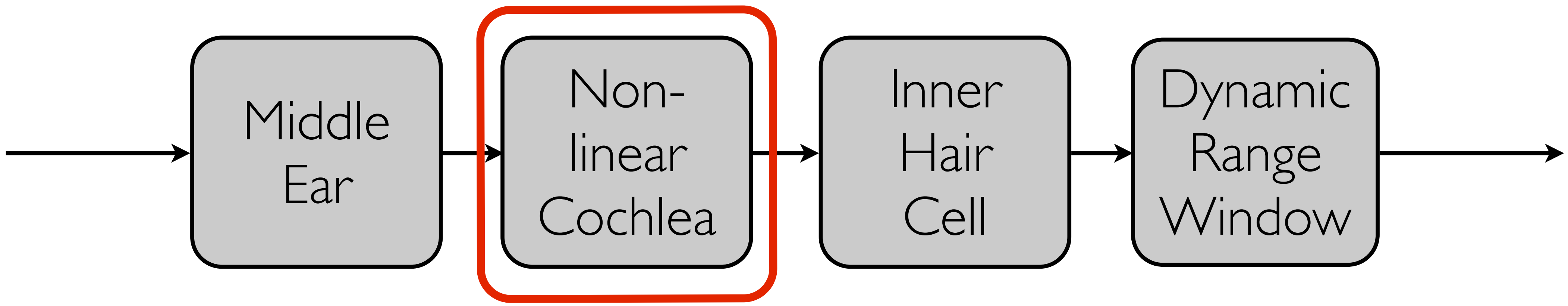


Model of Ascending Pathway



- **Middle Ear**
  - Modeled by a high-pass filter

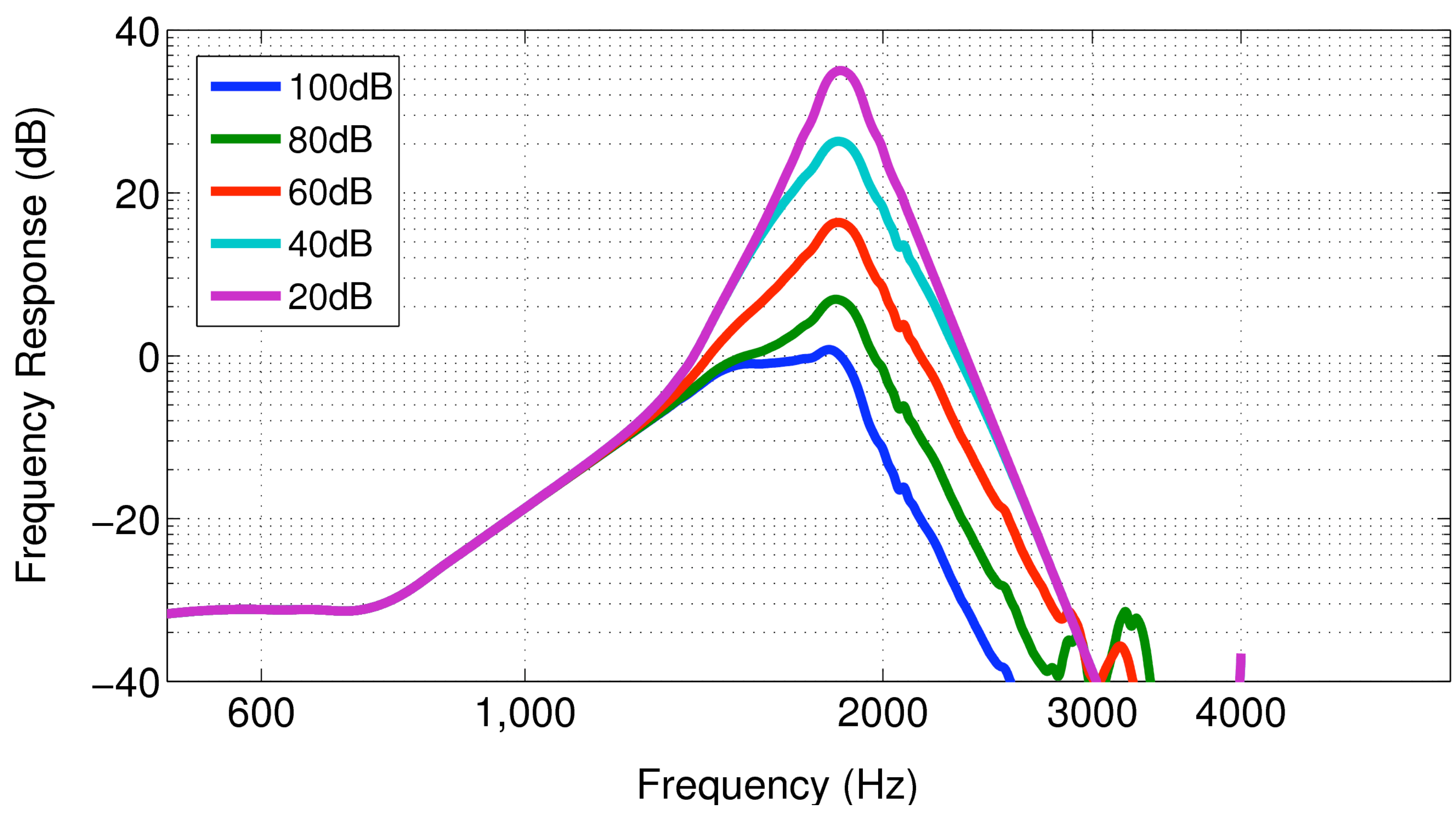
Model of Ascending Pathway



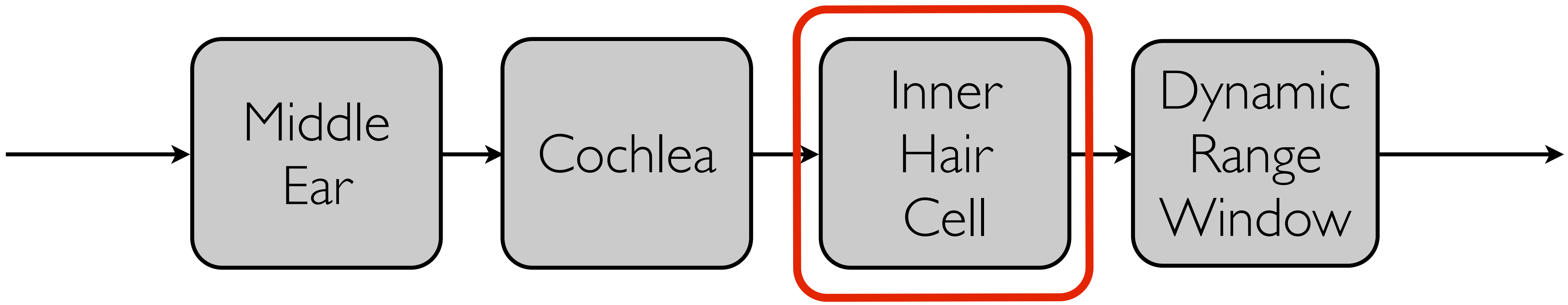
- J. Goldstein, 1990
- Multi-Band Path Non-Linear model (MBPNL)

MBPNL Model

- **Modeling cochlear nonlinearity**
- **Example for center frequency = 1820 Hz**
  - filter characteristics change <u>instantaneously</u> as a function of input signal strength

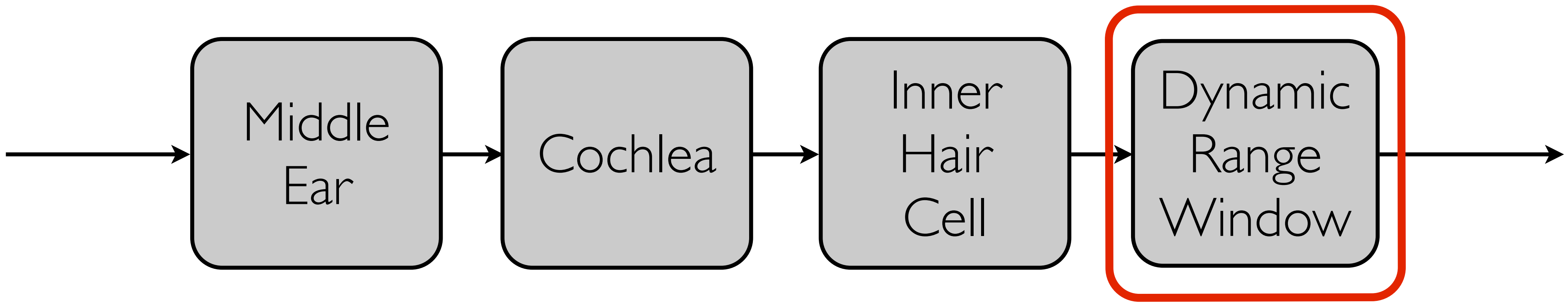


Model of Ascending Pathway



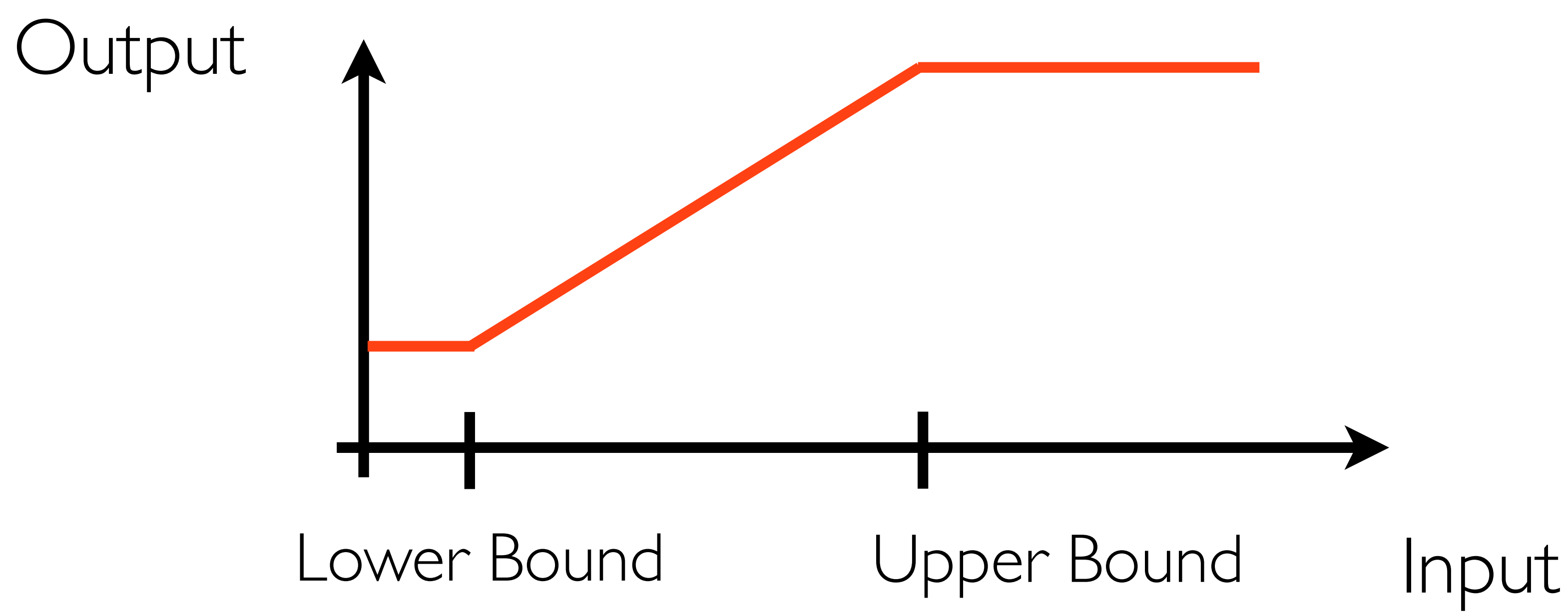
- **Inner Hair Cell**
  - Generic MIT model
  - A half-wave rectifier followed by a low pass filter

Model of Ascending Pathway



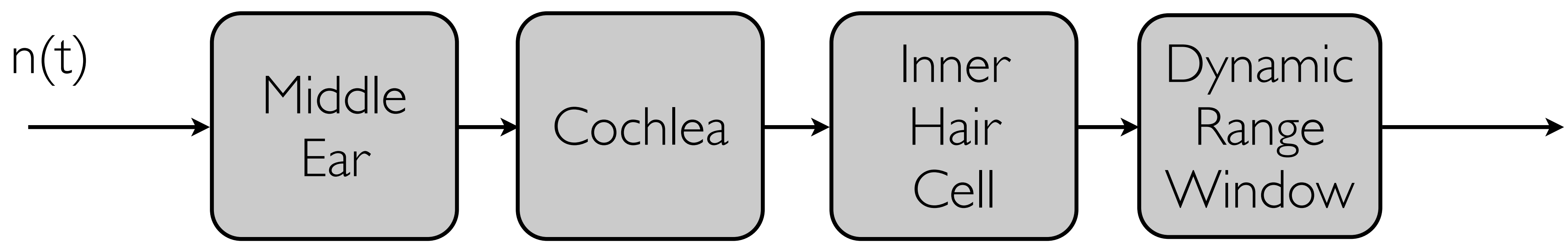
- **Dynamic Range Window (DRW)**
  - A hard limiter with upper and lower bounds, representing the dynamic range of auditory nerve firing

Dynamic Range Window

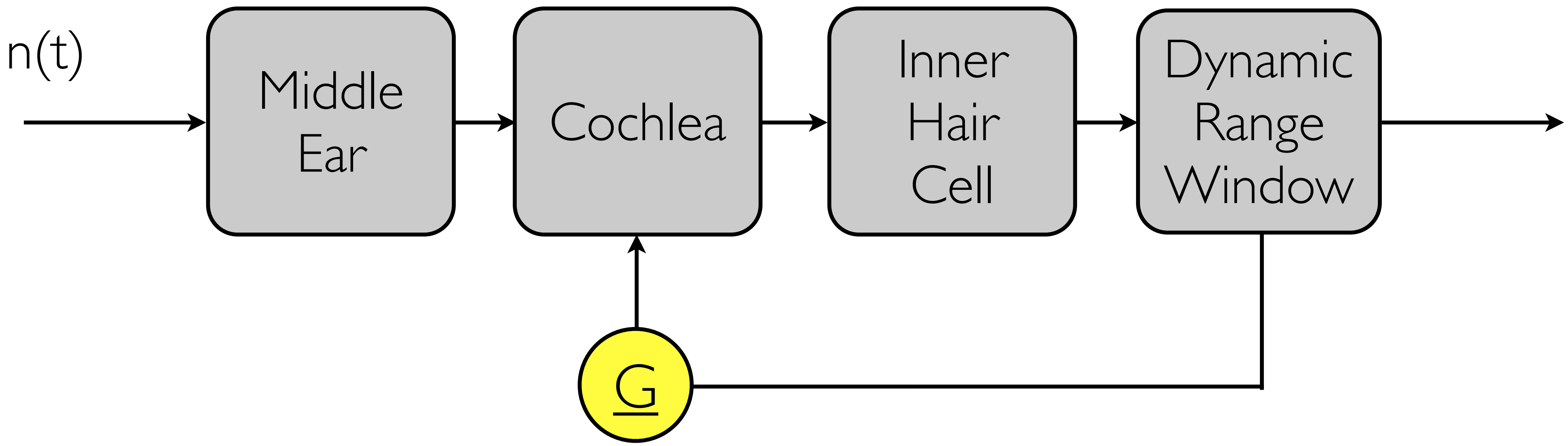


- No firing for signals below the lower bound
- Saturation in firing rate for signals above the upper bound

An Efferent-inspired Auditory Model

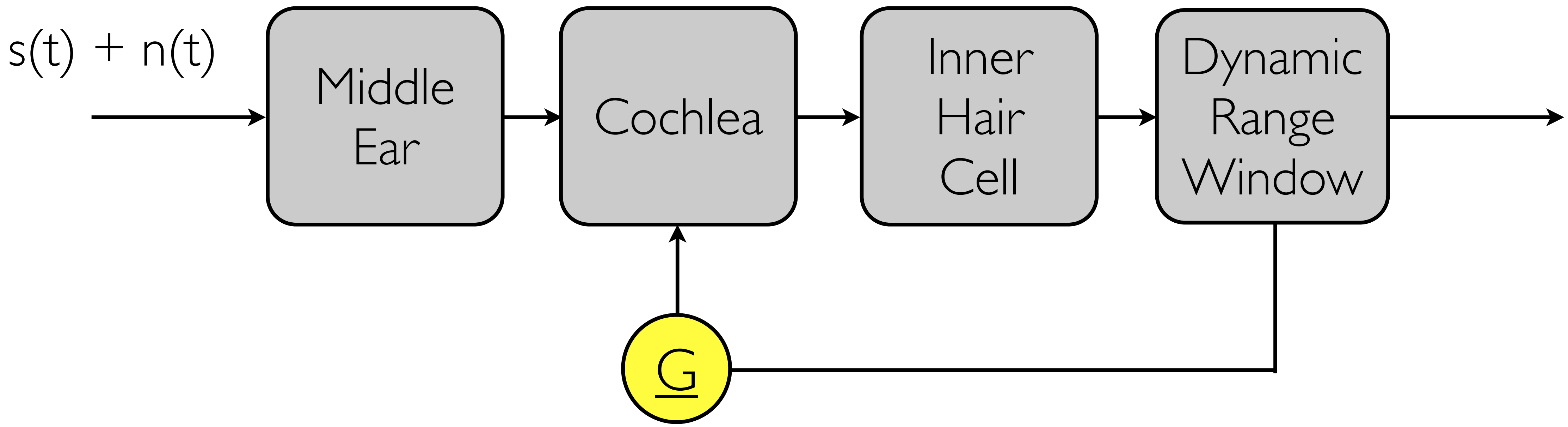


An Efferent-inspired Auditory Model



- G is adjusted based on the background noise such that the output of the DRW is at "epsilon level".
  - G impacts the filter response in the MBPNL cochlear model.

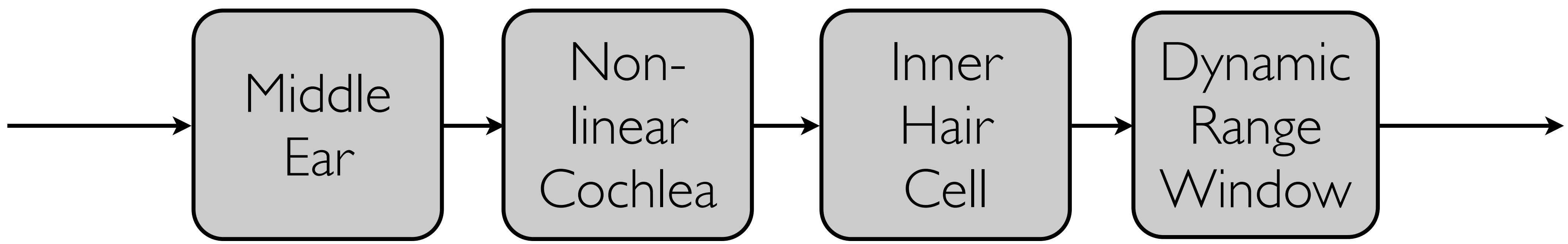
An Efferent-inspired Auditory Model



- The noisy speech signal is processed by the tuned auditory model.

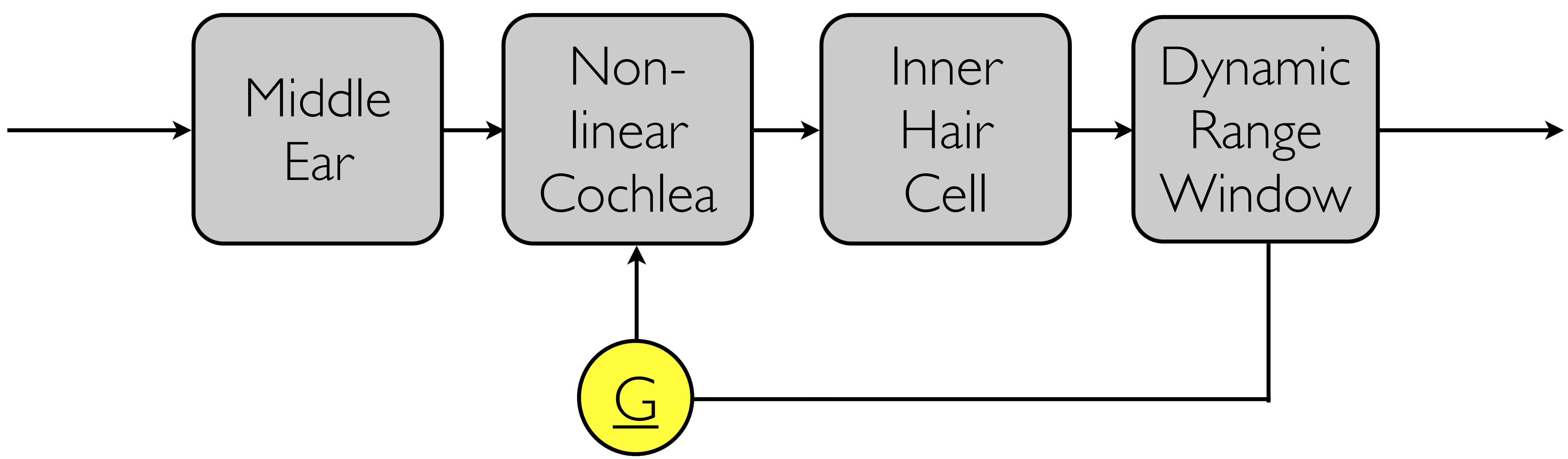
Definitions

- **Open-loop model**
  - The model for the ascending pathway



Definitions

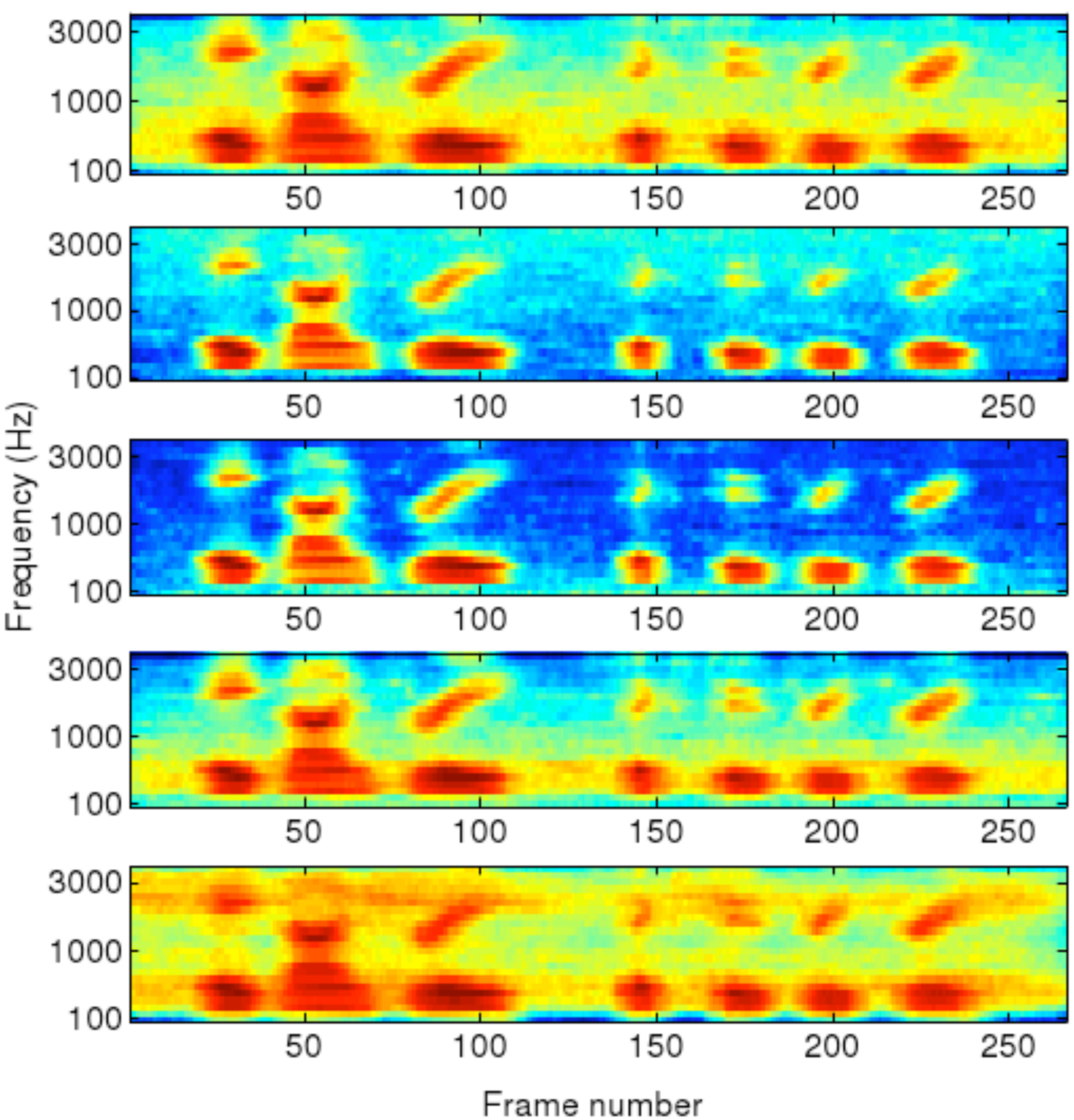
- **Closed-loop model**
  - The ascending pathway model with the efferent-inspired feedback



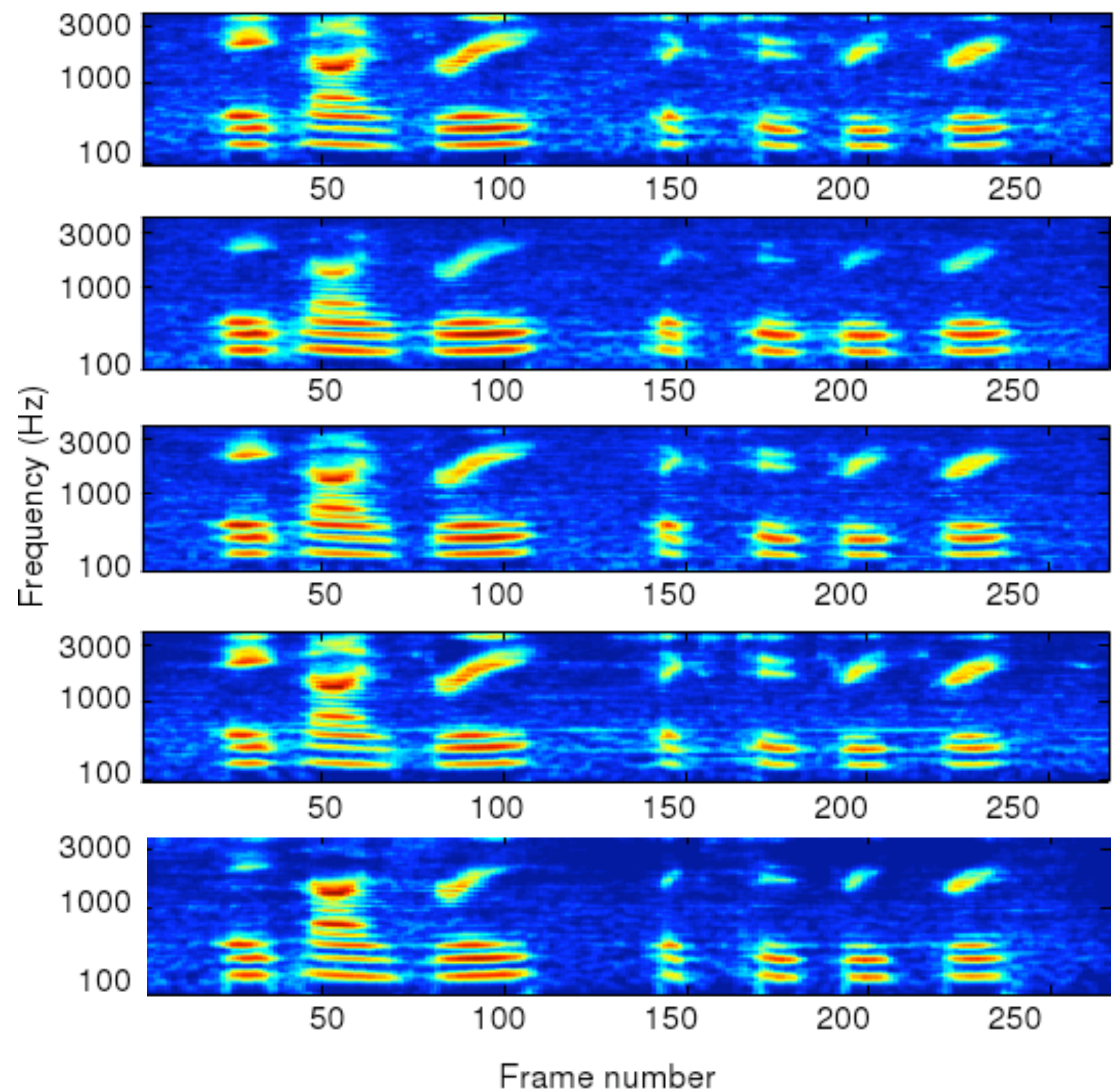
Visual Illustration

- Rows represent speech in different types of noise at 10 dB SNR

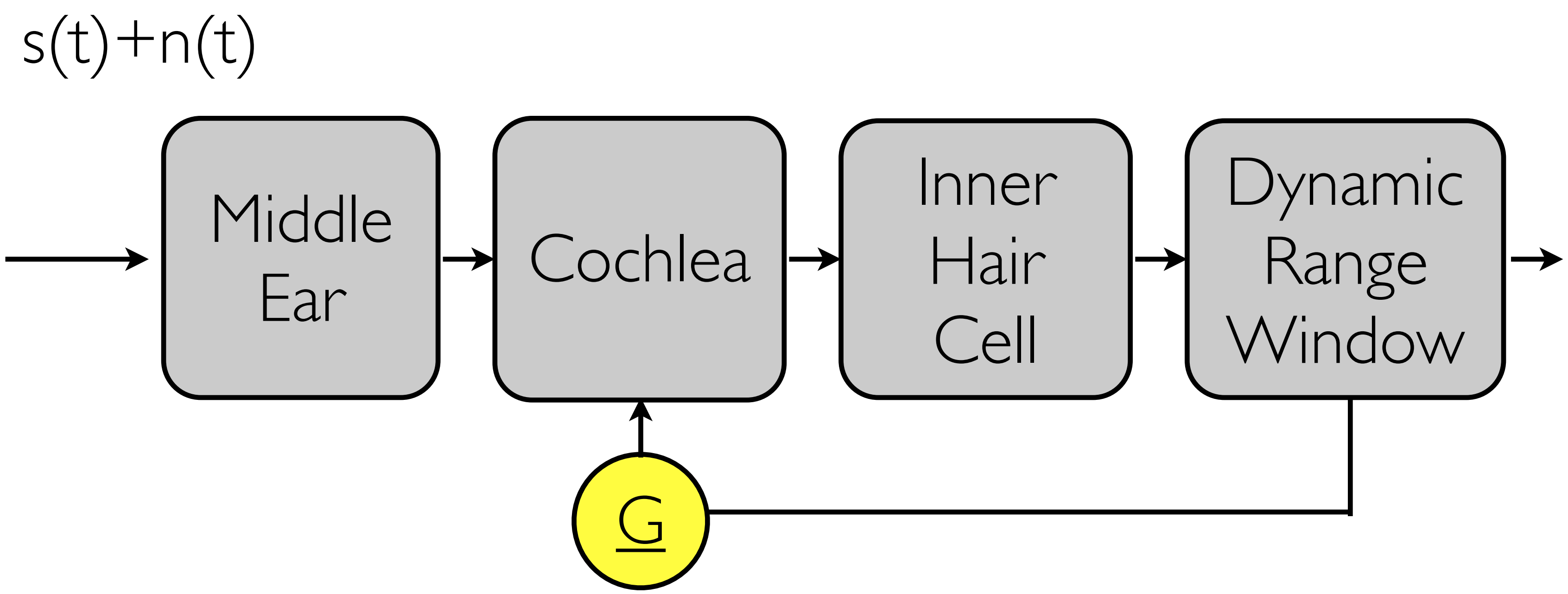
Short time Fourier transform



Closed-loop model

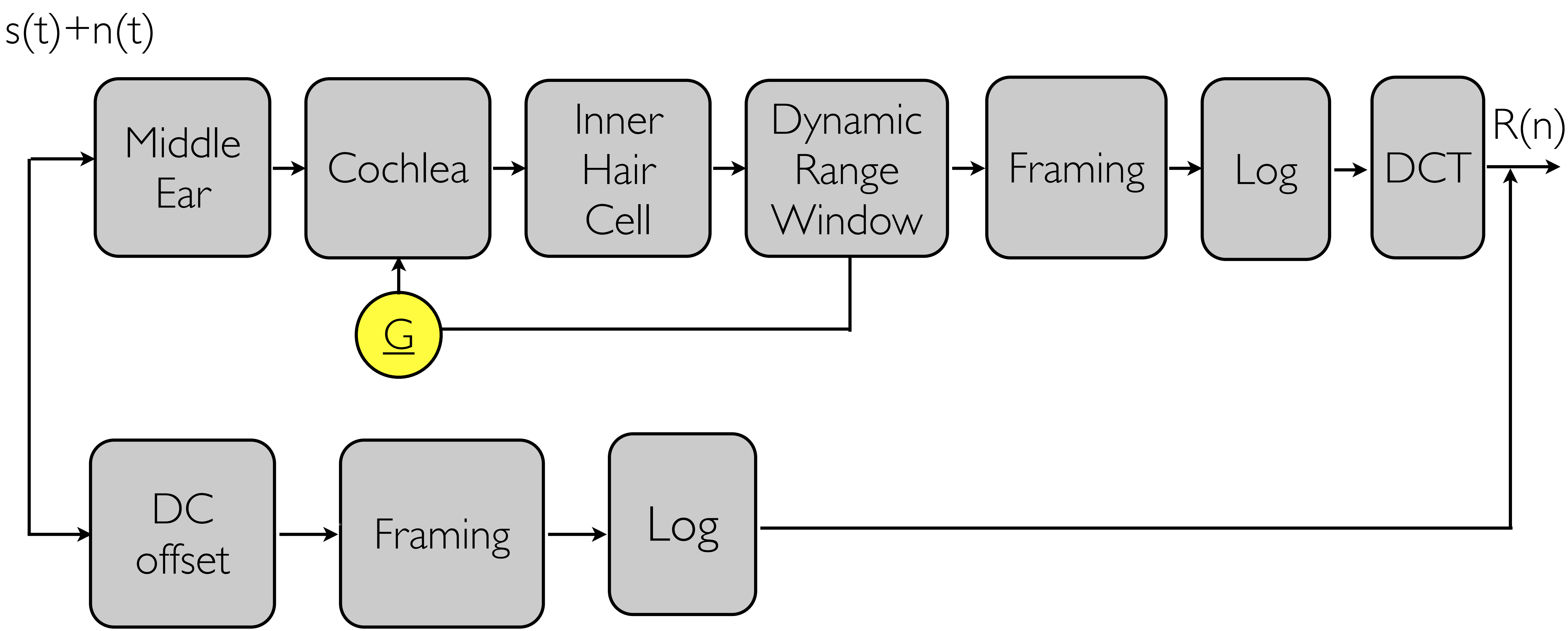


A Closed-loop Front-end for ASR



- Need to extract features that can be processed by speech recognizers

A Closed-loop Front-end for ASR



- The feature generation method follows the standard MFCC extraction process.

Experimental Setup

- Corpus creation (noisy speech data synthesis)
- Feature extraction methods
- Recognizer training and testing
- Experimental results

Corpus Creation

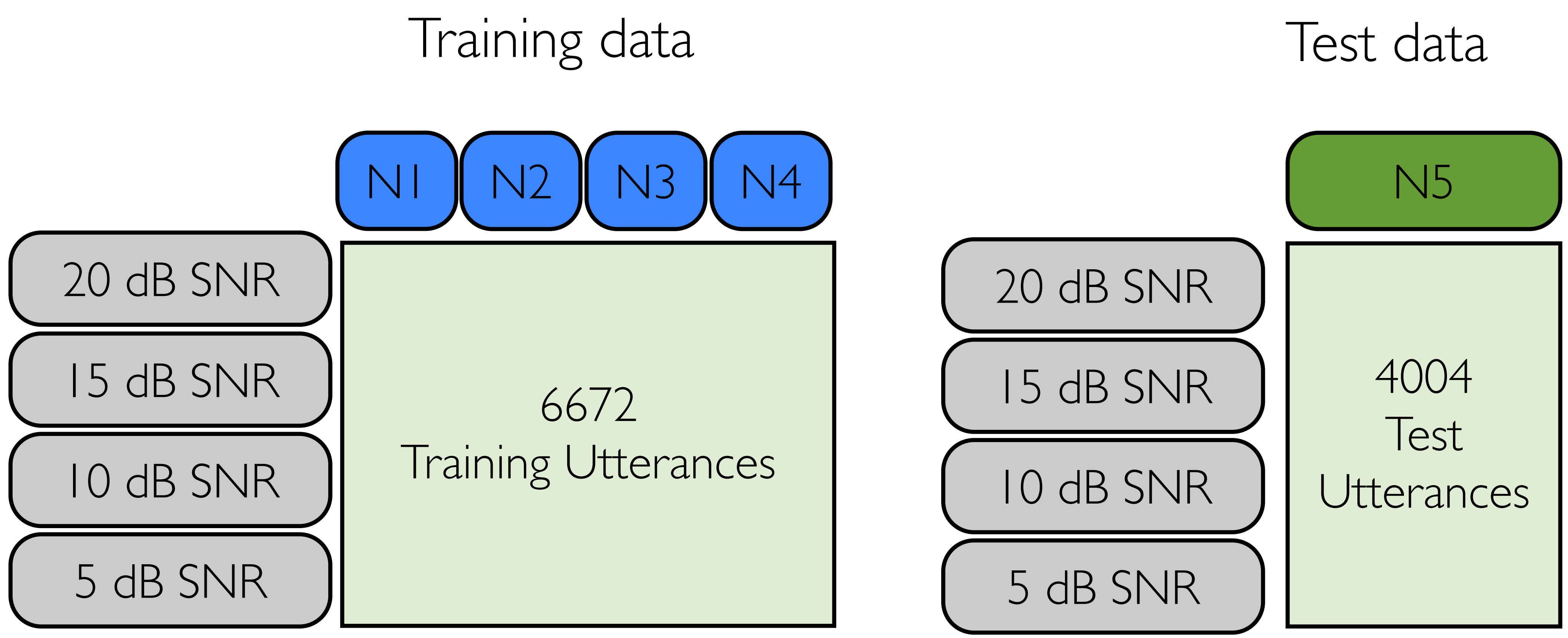
- **Noise signals**
  - Stationary noise: speech-shaped, white, pink
  - Non-stationary Aurora2 noise: train, subway
- **Speech signals**
  - Aurora2 digits (TIDigits)
- **Noisy speech synthesis**
  - Noise signals are fixed at 70 dB SPL
  - Speech signals are adjusted to create 5 to 20 dB SNRs
  - 300 ms adaptation prior to speech signal

Feature Extraction Methods

- **Three feature extraction methods**
  - MFCC baseline with conventional normalization method
  - The open-loop auditory model (in paper)
  - The closed-loop auditory model

Recognizer Training and Testing

- Standard Aurora2 HMM-based recognizer was used
- Jackknifing experiments with mismatched training and test conditions



Experimental Results

Accuracy (%)	MFCC Baseline	Closed-loop model
Average	86	92
STD	8.6	4.7

- The closed-loop model performs 43% better than the MFCC baseline, and reduced variation across mismatched conditions by 45%.

Experimental Results

MFCC baseline

Acc (%) / dB SNR	speech-shaped	White	Pink	Subway	Train
20	95	92	91	88	94
15	94	90	89	84	93
10	91	85	85	76	92
5	81	73	76	62	84
Avg	90	85	85	77	91

Closed-loop model

Acc (%) / dB SNR	speech-shaped	White	Pink	Subway	Train
20	96	94	95	93	96
15	96	93	96	92	95
10	94	91	95	89	93
5	83	83	91	78	84
Avg	92	90	94	88	92

- The closed-loop model performed better than the baseline across all mismatched training and test conditions.

Conclusions

- **Key ideas**
  - Efferent-inspired feedback regulates the operating point of the front-end
  - Results in a stable representation -- a desired property for ASR
- **Experimental validation**
  - Digit recognition in noise in mismatched conditions with multiple noise types and SNRs
  - The closed-loop model outperformed the baseline across all mismatched training and test conditions.
  - The results indicate that incorporating feedback in the front-end shows promise for generating robust speech features.