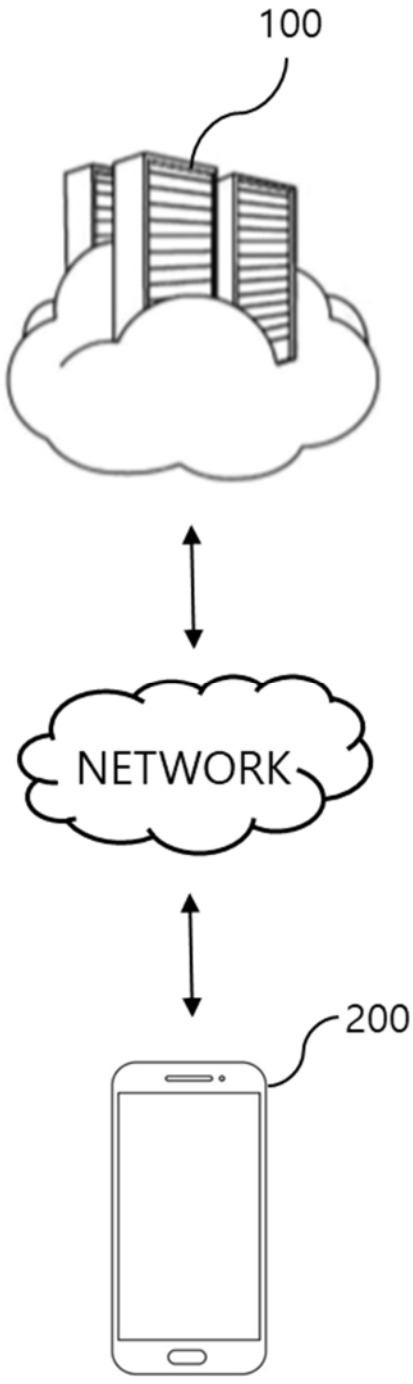
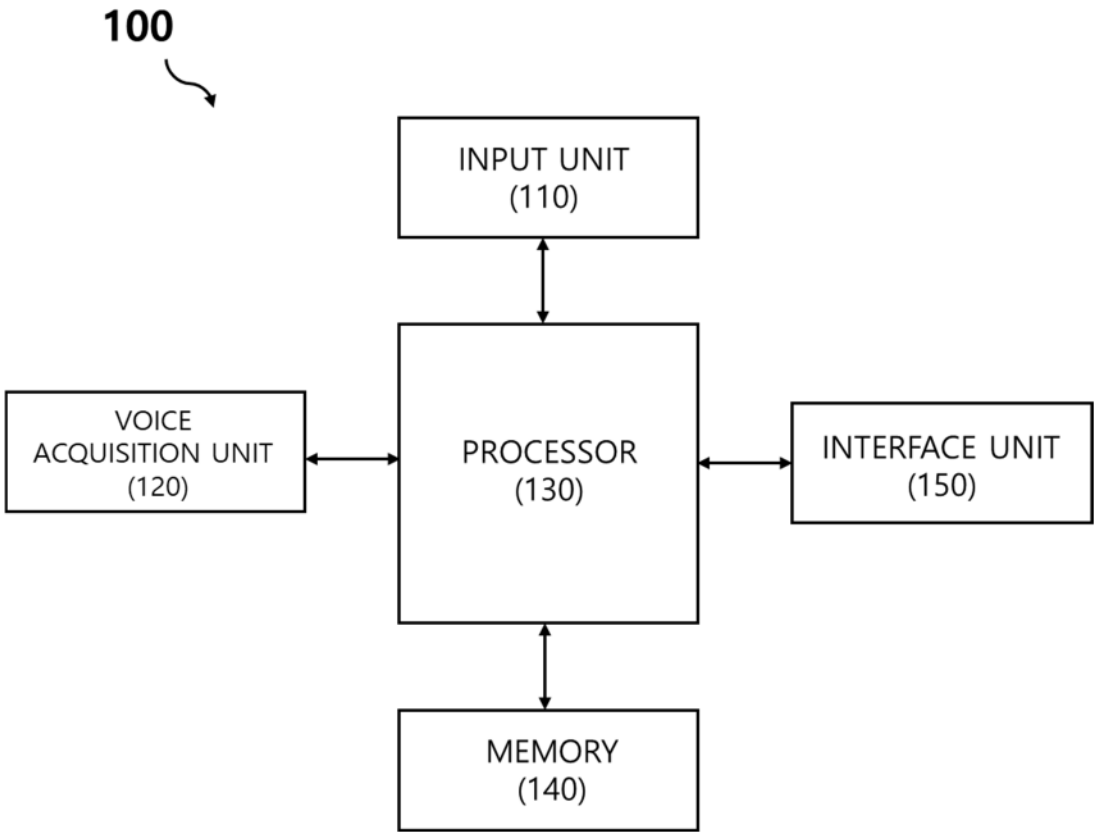
[Drawings]

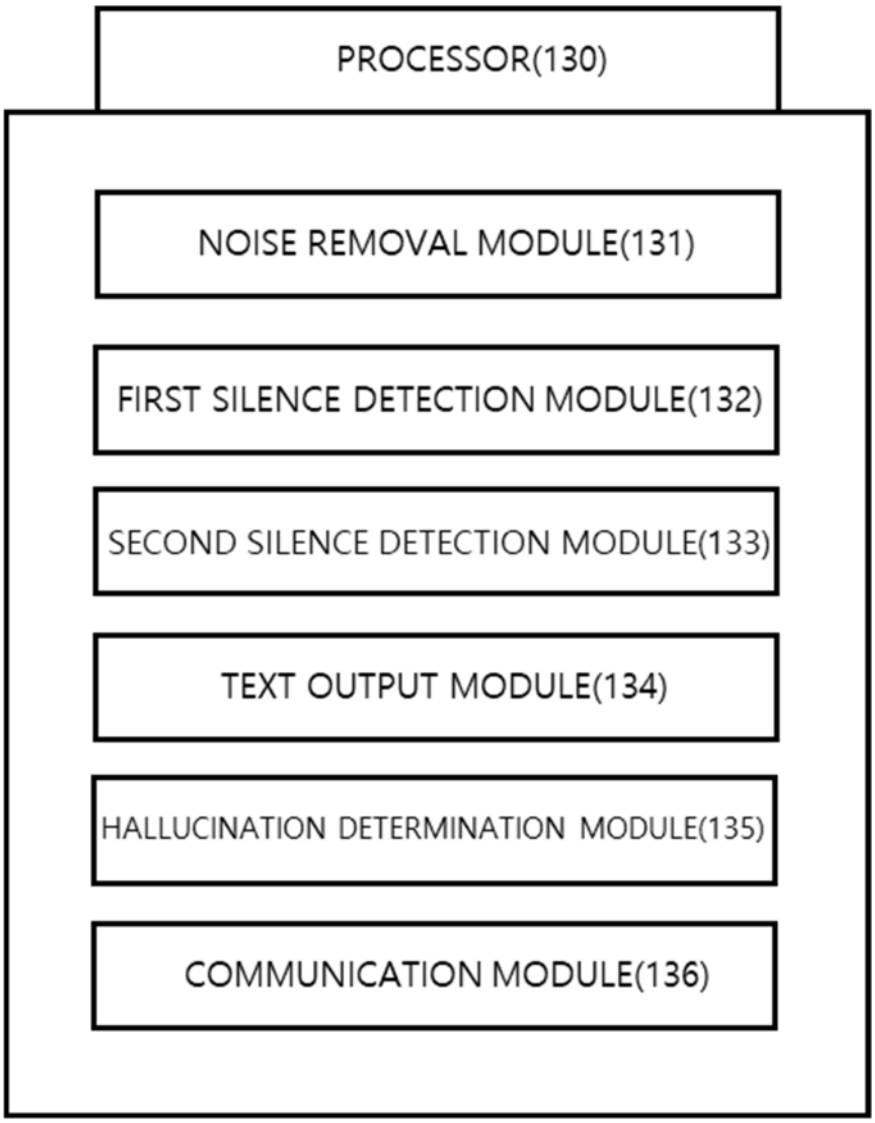
[FIG. 1]



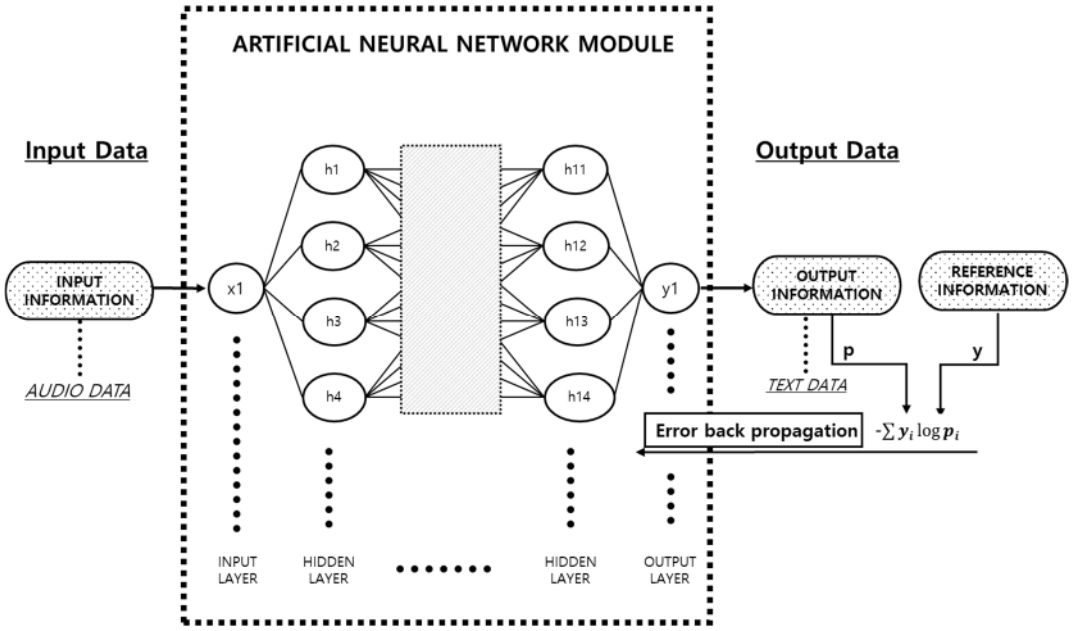
[FIG. 2]


100
INPUT UNIT
(110)
VOICE
ACQUISITION UNIT
(120)
PROCESSOR
(130)
INTERFACE UNIT
(150)
MEMORY
(140)

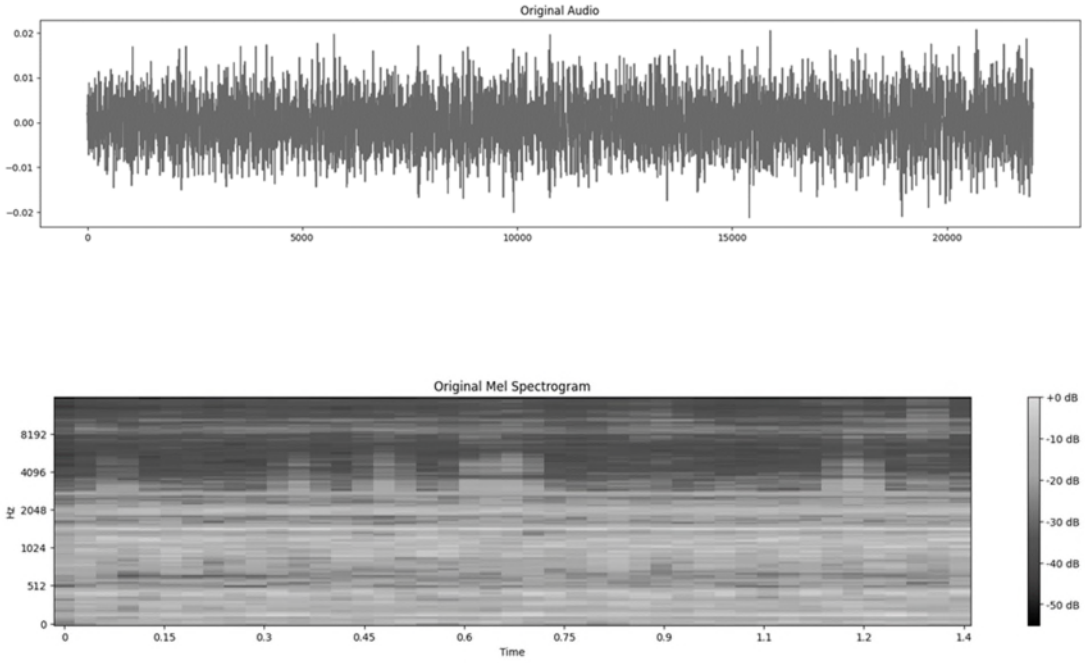
[FIG. 3]


PROCESSOR(130)
NOISE REMOVAL MODULE(131)
FIRST SILENCE DETECTION MODULE(132)
SECOND SILENCE DETECTION MODULE(133)
TEXT OUTPUT MODULE(134)
HALLUCINATION DETERMINATION MODULE(135)
COMMUNICATION MODULE(136)

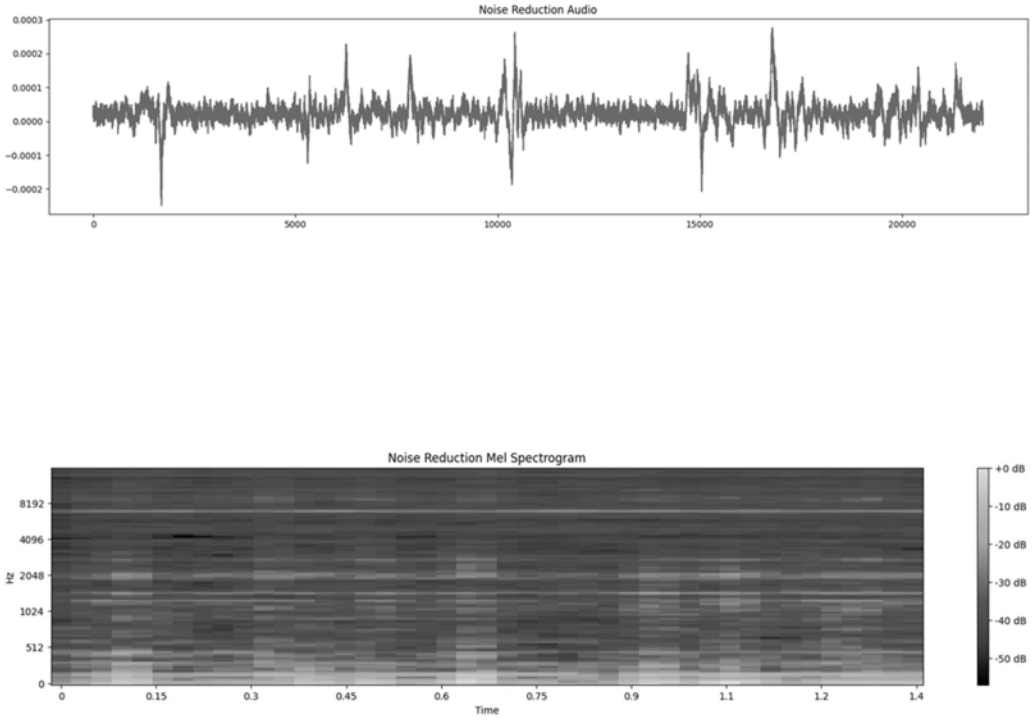
[FIG. 4]


ARTIFICIAL NEURAL NETWORK MODULE
Input Data
Output Data
INPUT INFORMATION
AUDIO DATA
x1
h1
h2
h3
h4
h11
h12
h13
h14
y1
OUTPUT INFORMATION
REFERENCE INFORMATION
TEXT DATA
p
y
Error back propagation
$-\sum y_i \log p_i$
INPUT LAYER
HIDDEN LAYER
HIDDEN LAYER
OUTPUT LAYER

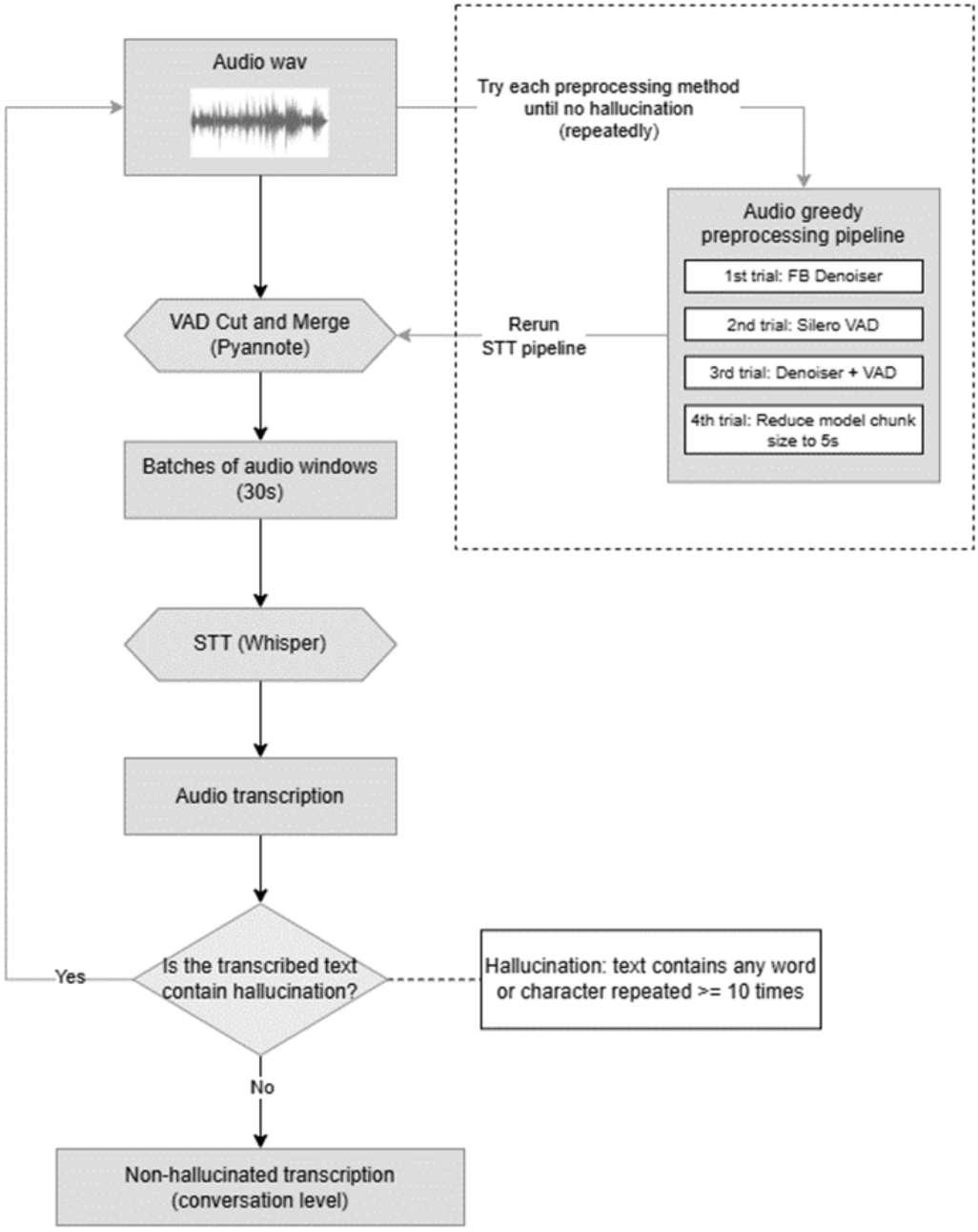
[FIG. 5]


Original Audio
0.02
0.01
0.00
−0.01
−0.02
0
5000
10000
15000
20000
Original Mel Spectrogram
Hz
8192
4096
2048
1024
512
0
0
0.15
0.3
0.45
0.6
0.75
0.9
1.1
1.2
1.4
Time
+0 dB
-10 dB
-20 dB
-30 dB
-40 dB
-50 dB

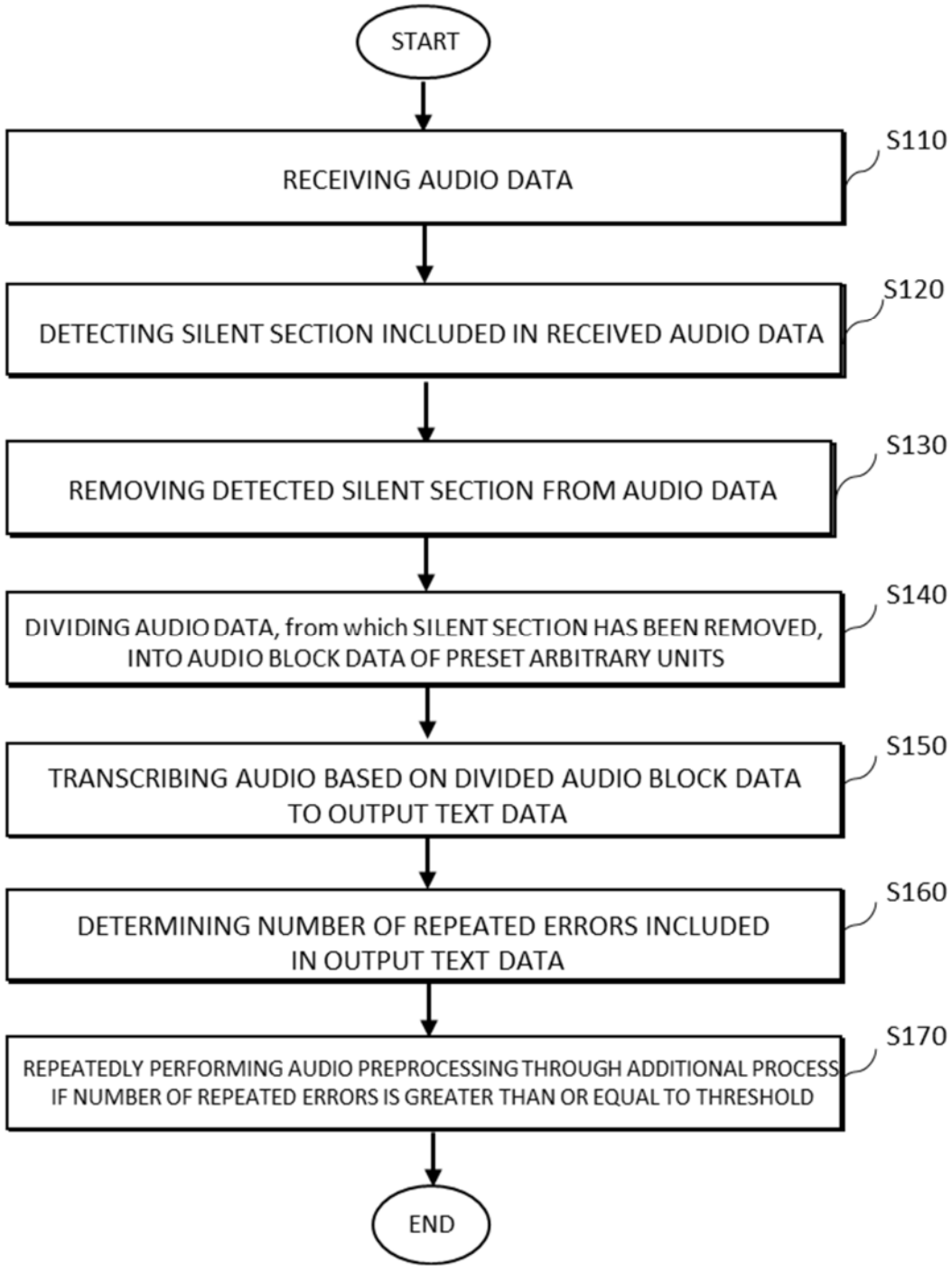
[FIG. 6]


Noise Reduction Audio
0.0003
0.0002
0.0001
0.0000
−0.0001
−0.0002
0
5000
10000
15000
20000
Noise Reduction Mel Spectrogram
Hz
8192
4096
2048
1024
512
0
0
0.15
0.3
0.45
0.6
0.75
0.9
1.1
1.2
1.4
Time
+0 dB
-10 dB
-20 dB
-30 dB
-40 dB
-50 dB

[FIG. 7]


Audio wav
Try each preprocessing method until no hallucination (repeatedly)
Audio greedy preprocessing pipeline
1st trial: FB Denoiser
2nd trial: Silero VAD
3rd trial: Denoiser + VAD
4th trial: Reduce model chunk size to 5s
Rerun STT pipeline
VAD Cut and Merge (Pyannote)
Batches of audio windows (30s)
STT (Whisper)
Audio transcription
Is the transcribed text contain hallucination?
Yes
Hallucination: text contains any word or character repeated >= 10 times
No
Non-hallucinated transcription (conversation level)

[FIG. 8]


START
RECEIVING AUDIO DATA
S110
DETECTING SILENT SECTION INCLUDED IN RECEIVED AUDIO DATA
S120
REMOVING DETECTED SILENT SECTION FROM AUDIO DATA
S130
DIVIDING AUDIO DATA, from which SILENT SECTION HAS BEEN REMOVED, INTO AUDIO BLOCK DATA OF PRESET ARBITRARY UNITS
S140
TRANSCRIBING AUDIO BASED ON DIVIDED AUDIO BLOCK DATA TO OUTPUT TEXT DATA
S150
DETERMINING NUMBER OF REPEATED ERRORS INCLUDED IN OUTPUT TEXT DATA
S160
REPEATEDLY PERFORMING AUDIO PREPROCESSING THROUGH ADDITIONAL PROCESS IF NUMBER OF REPEATED ERRORS IS GREATER THAN OR EQUAL TO THRESHOLD
S170
END