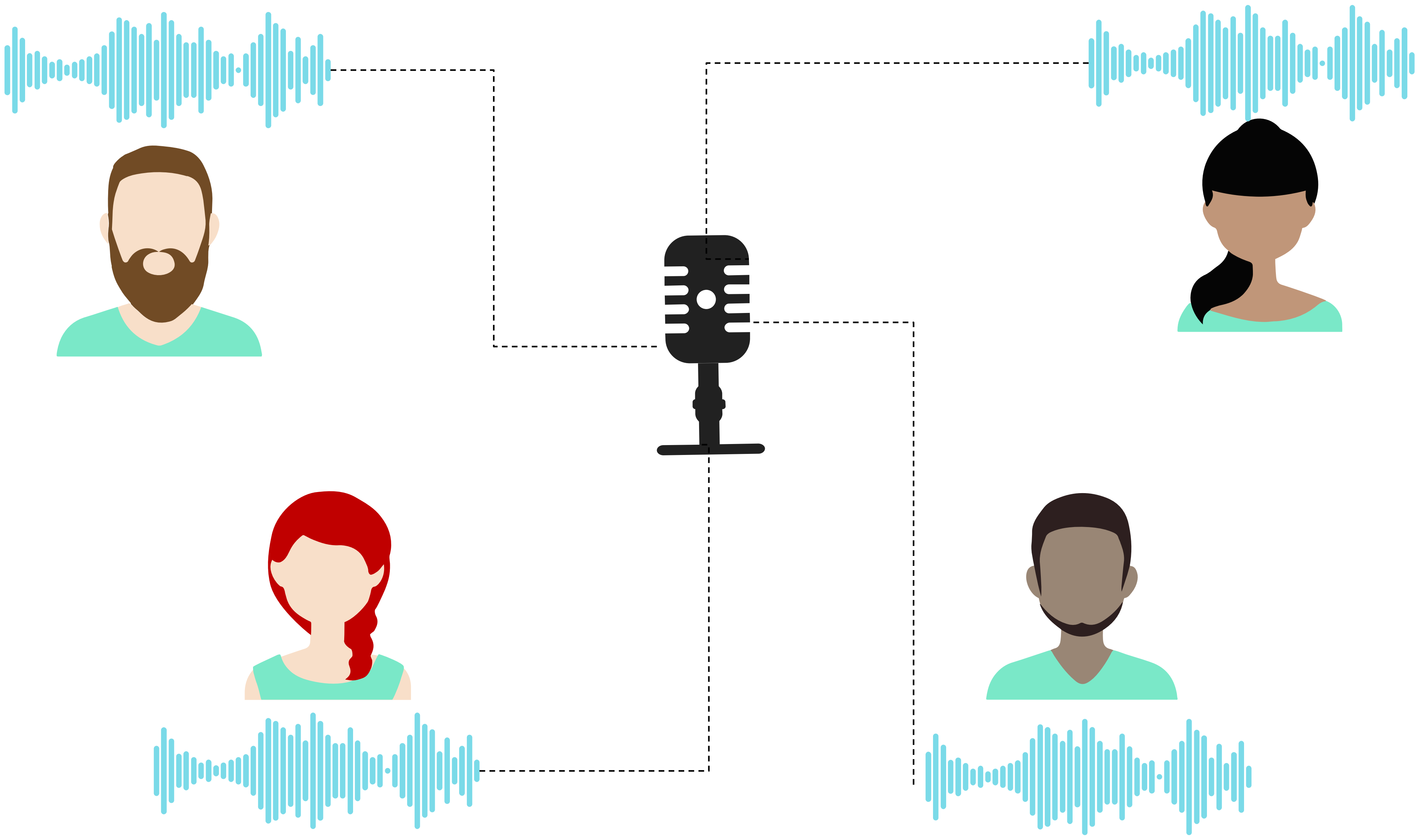
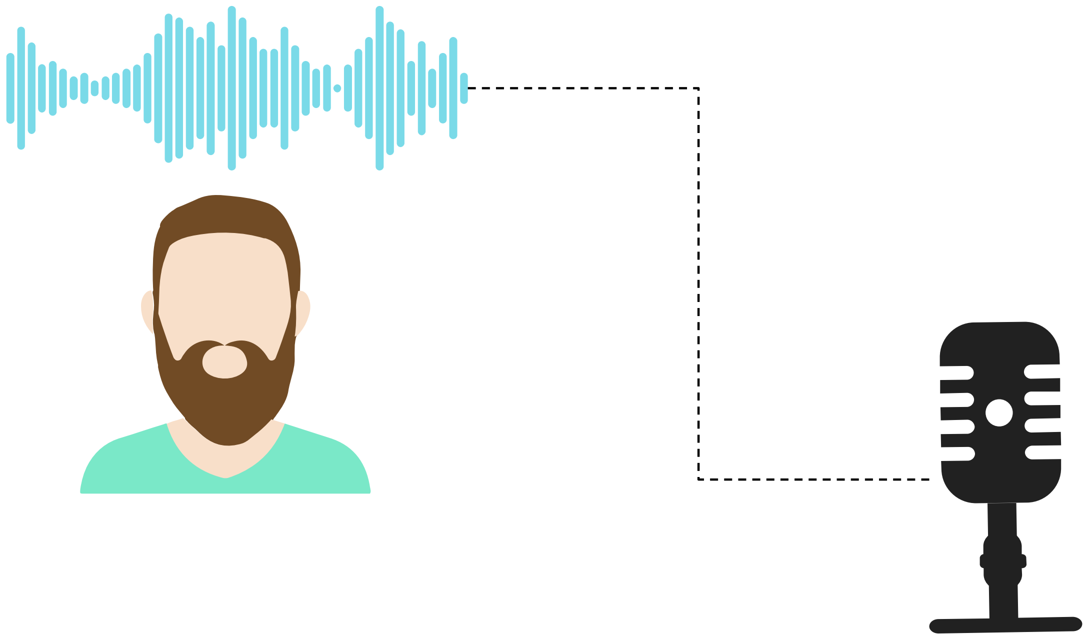
VOICEFILTER

TARGETED VOICE SEPARATION BY SPEAKER-CONDITIONED SPECTROGRAM MASKING

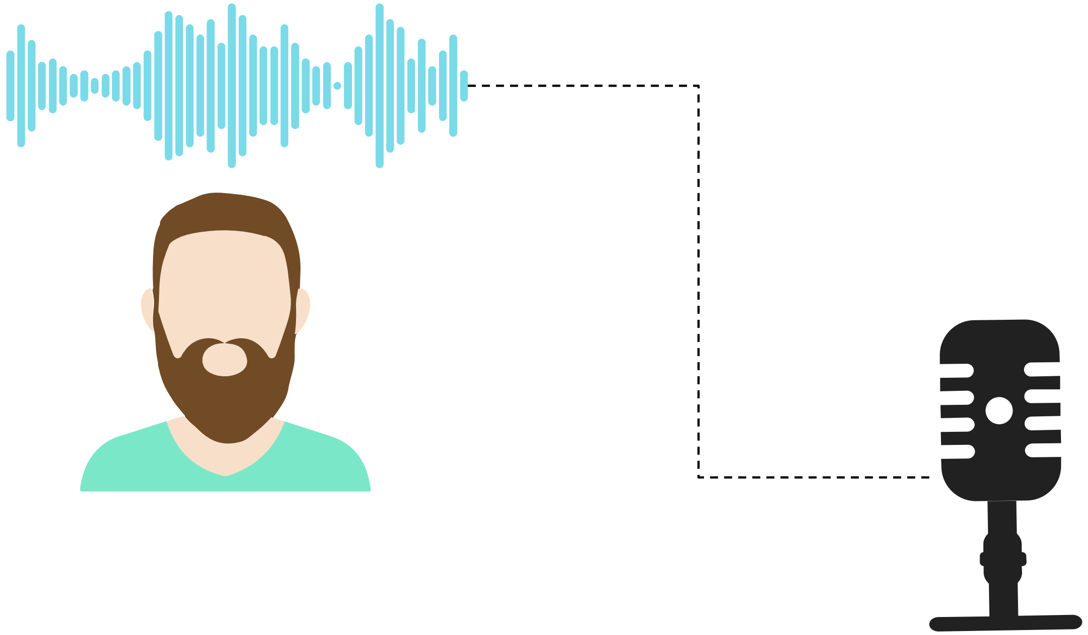
COCKTAIL PARTY PROBLEM



COCKTAIL PARTY PROBLEM



COCKTAIL PARTY PROBLEM



MULTI-CHANNEL

BLIND SEPARATION

SINGLE CHANNEL

BLIND SEPARATION

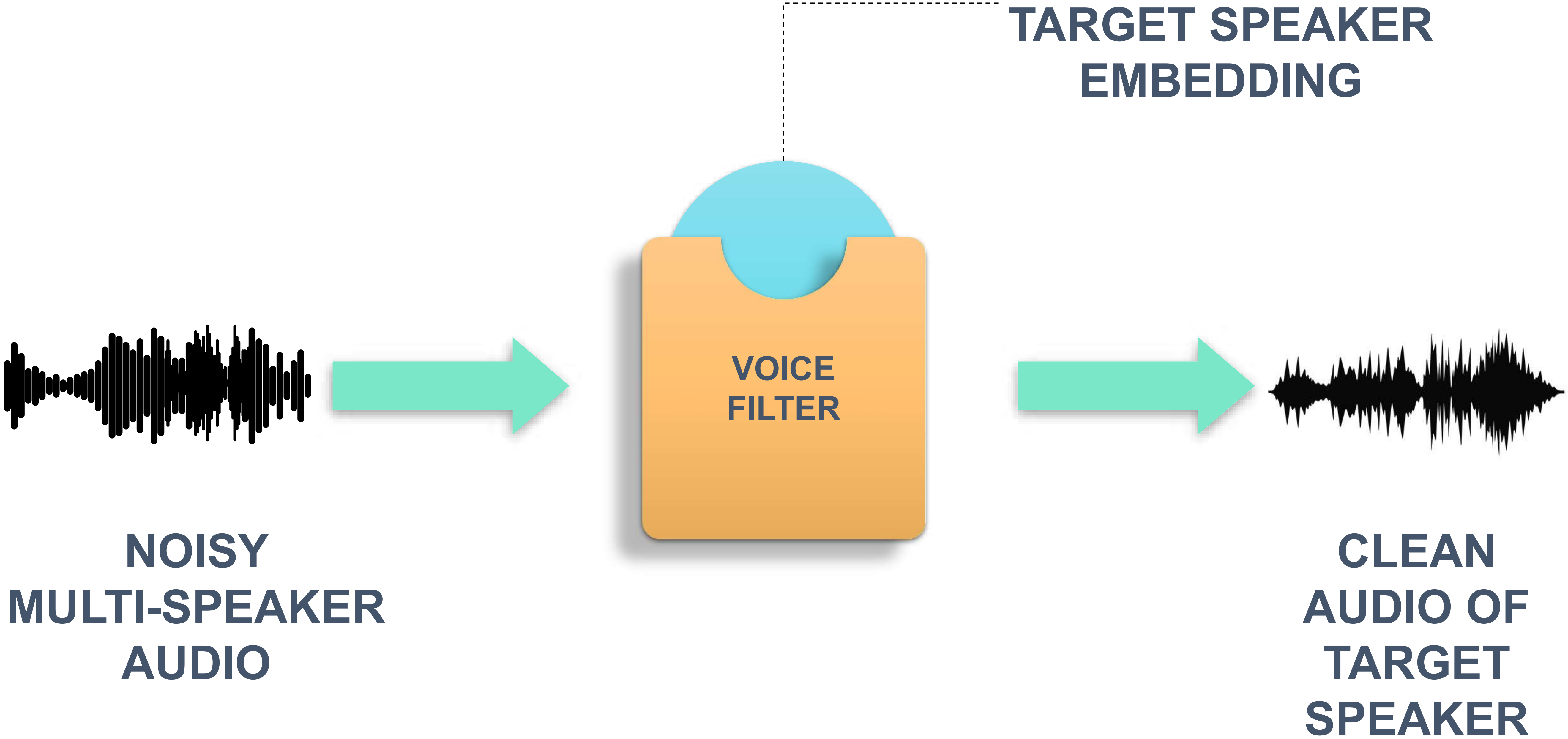
SINGLE CHANNEL BLIND SEPARATION SOLUTIONS

1. DEEP CLUSTERING, 2016 MITSUBISHI ELECTRIC RESEARCH LABORATORIES
2. DEEP ATTRACTOR NETWORK, 2017 COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK
3. PERMUTATION INVARIANT TRAINING, 2017 Microsoft Research

WE USUALLY KNOW WHOM TO LISTEN TO

WHOM TO LISTEN TO

- STORES A VOICE PROFILE
- WHICH MEANS THAT WE ARE NO LONGER BLIND



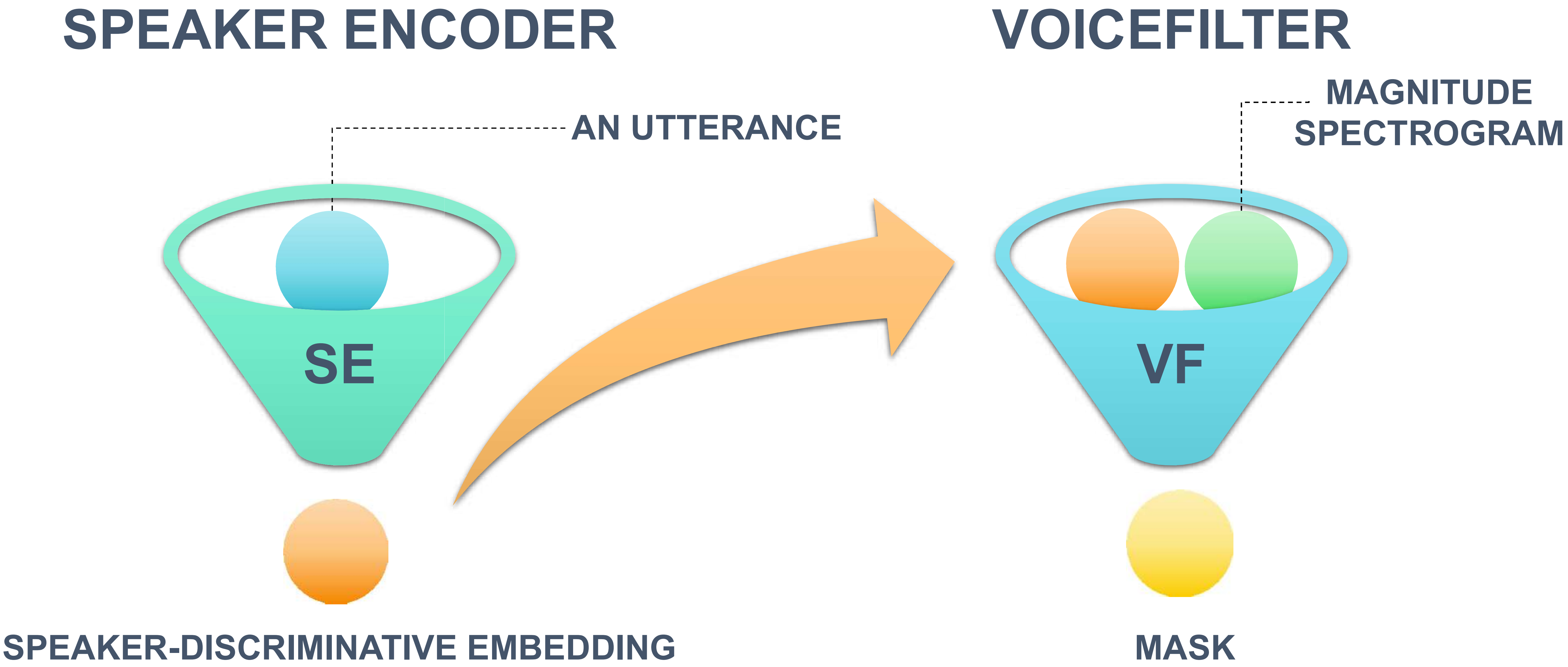
TARGET SPEAKER EMBEDDING

NOISY MULTI-SPEAKER AUDIO

VOICE FILTER

CLEAN AUDIO OF TARGET SPEAKER

MODELS



SPEAKER ENCODER

VOICEFILTER

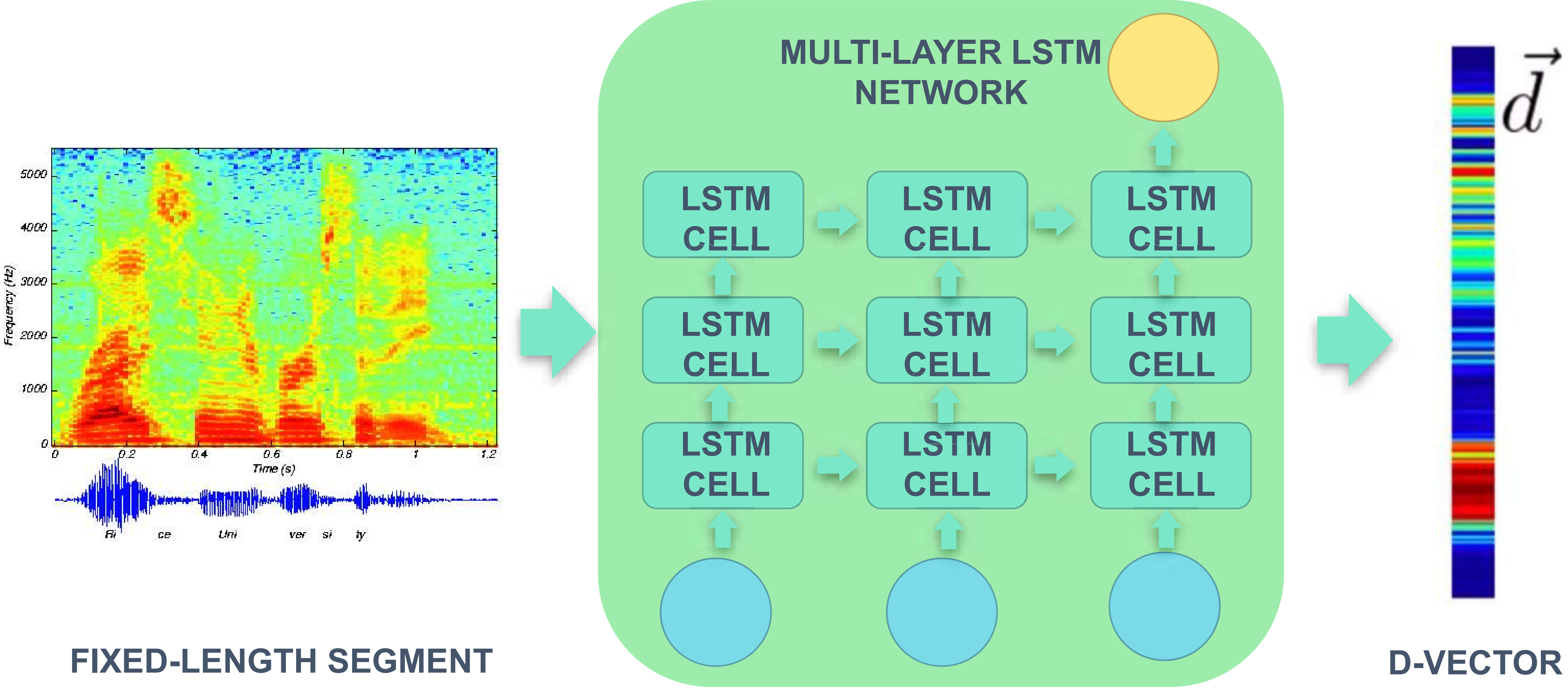
AN UTTERANCE

MAGNITUDE SPECTROGRAM

SPEAKER-DISCRIMINATIVE EMBEDDING

MASK

SPEAKER ENCODER
(THE D-VECTOR SYSTEM)



FIXED-LENGTH SEGMENT

MULTI-LAYER LSTM NETWORK

LSTM CELL

LSTM CELL

LSTM CELL

LSTM CELL

LSTM CELL

LSTM CELL

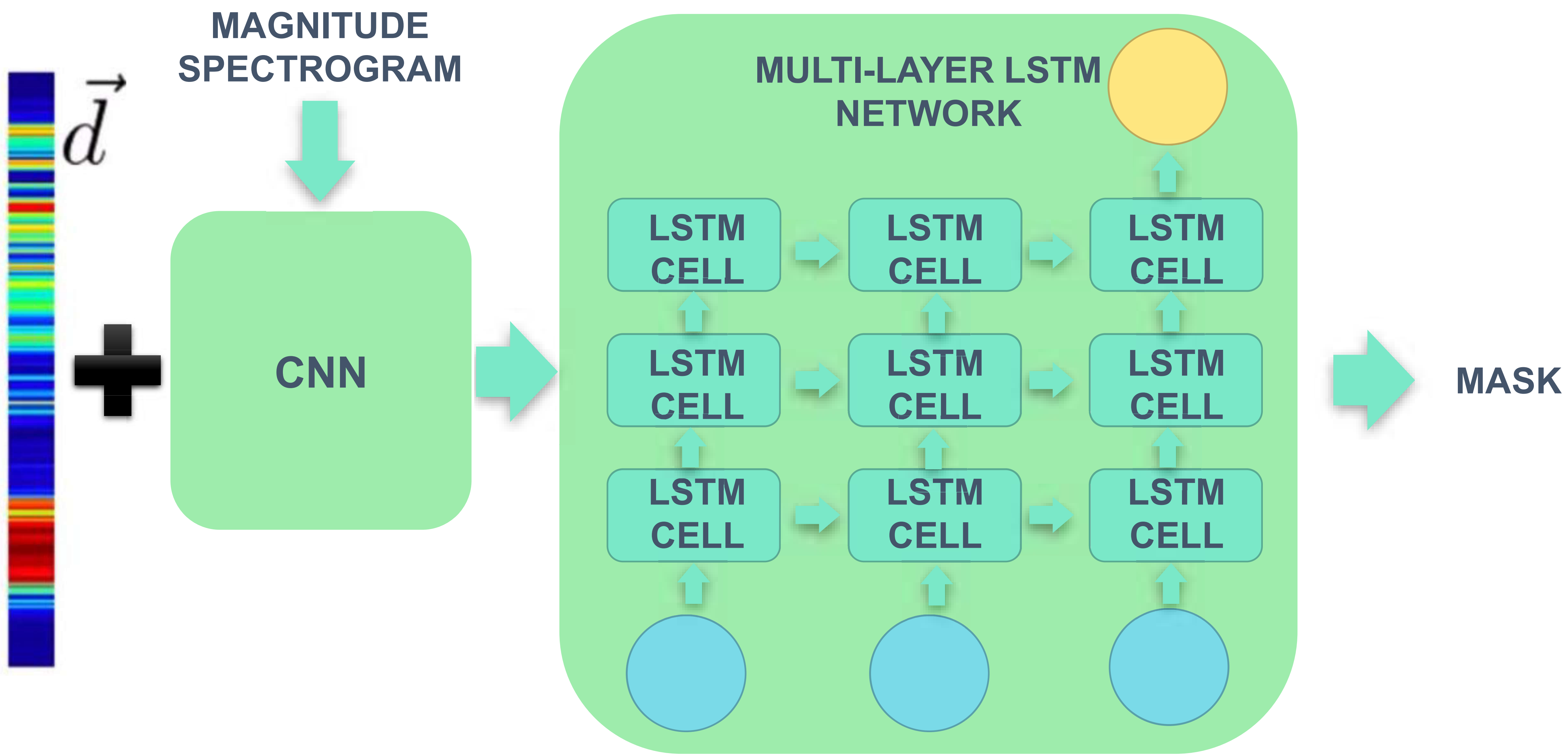
LSTM CELL

LSTM CELL

LSTM CELL

D-VECTOR

VOICEFILTER



$\vec{d}$

MAGNITUDE SPECTROGRAM

CNN

MULTI-LAYER LSTM NETWORK

LSTM CELL

LSTM CELL

LSTM CELL

LSTM CELL

LSTM CELL

LSTM CELL

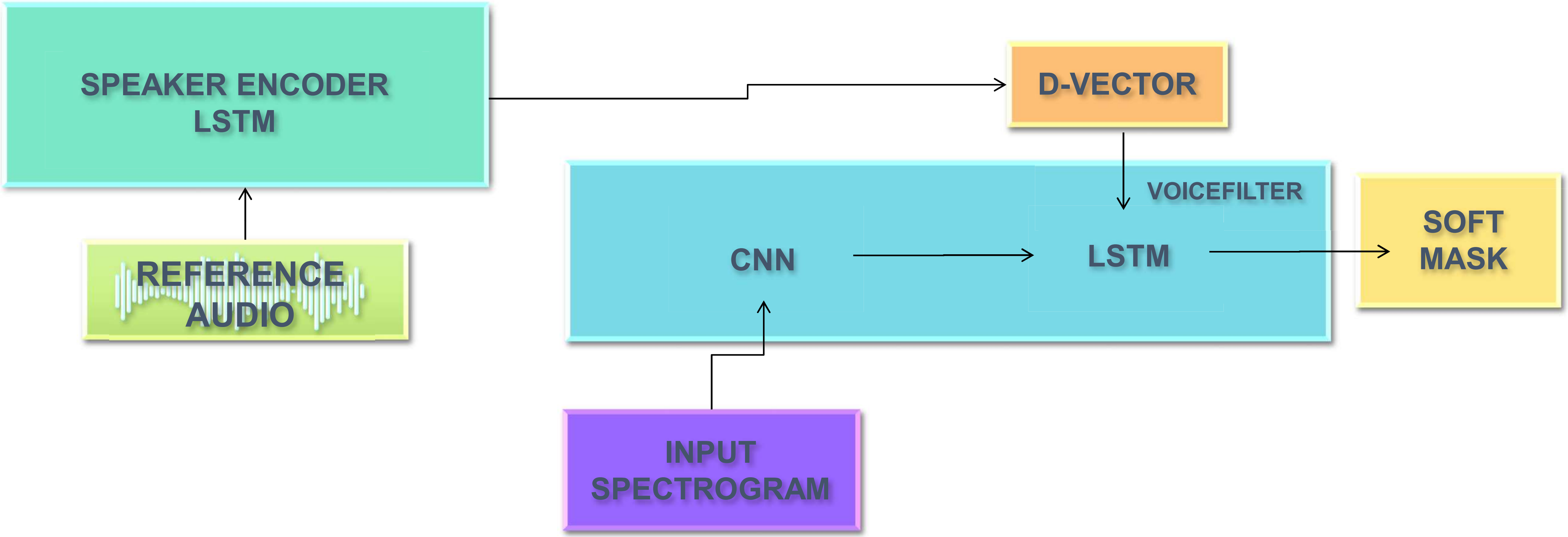
LSTM CELL

LSTM CELL

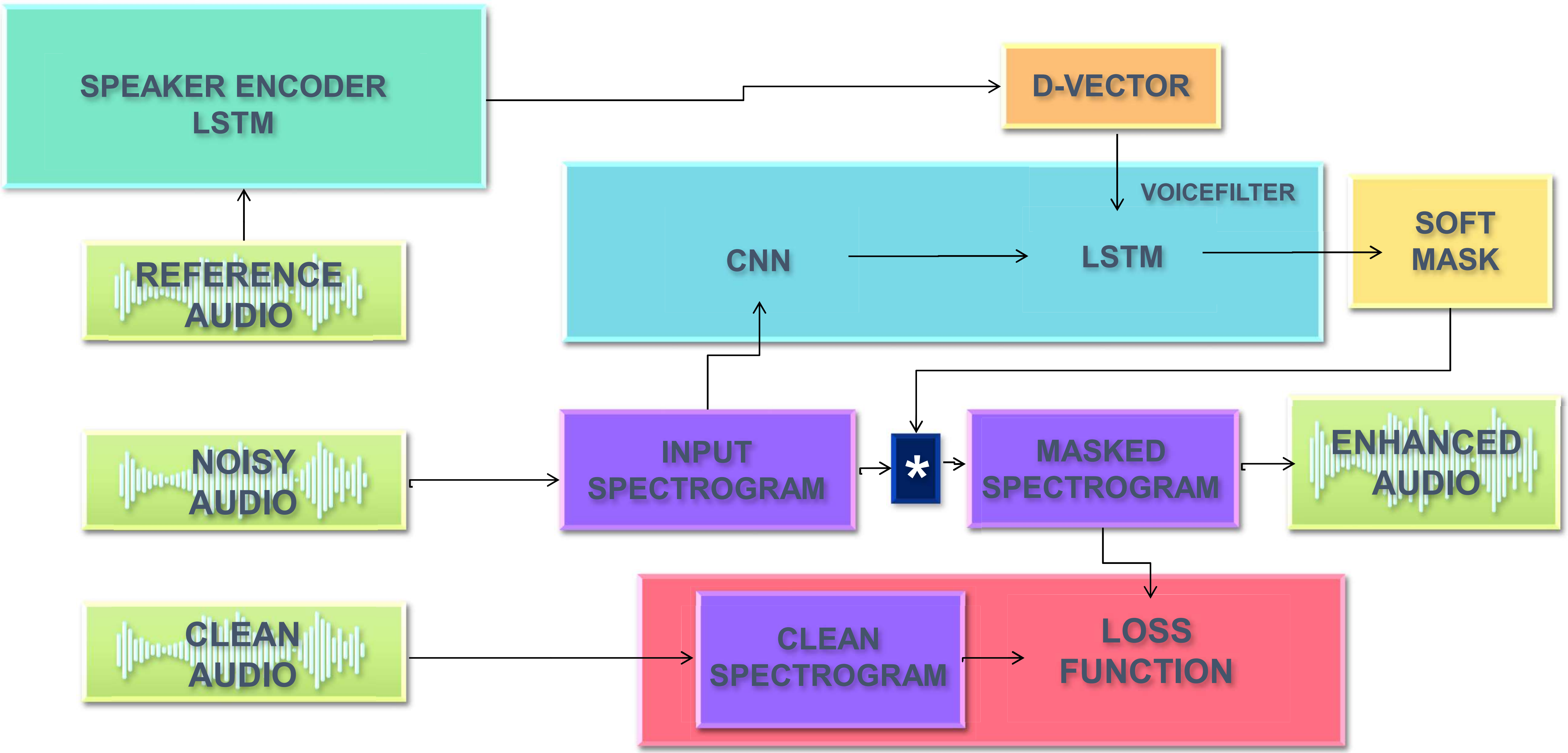
LSTM CELL

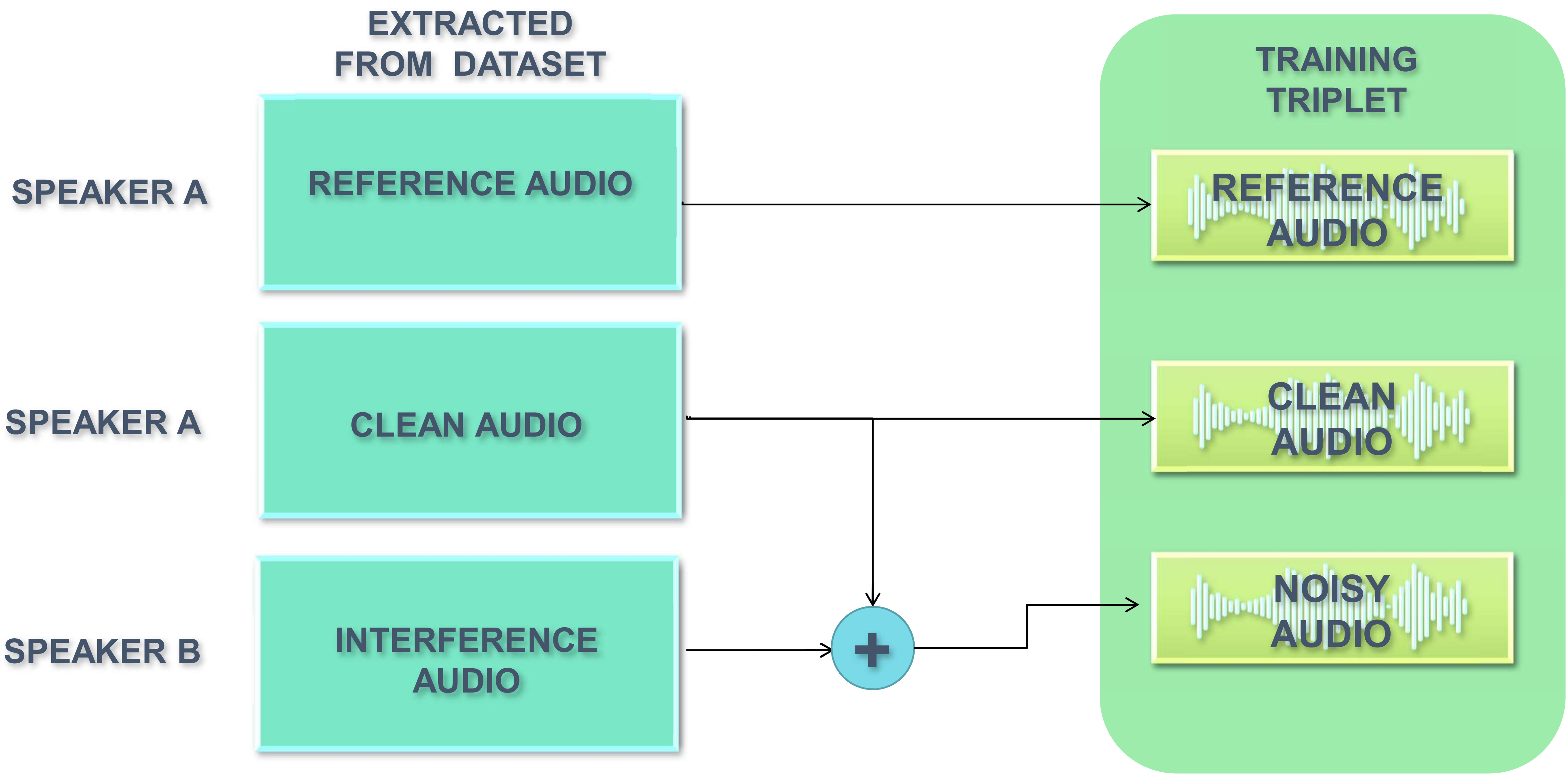
MASK

SYSTEM ARCHITECTURE



SYSTEM ARCHITECTURE




EXTRACTED
FROM DATASET
SPEAKER A
REFERENCE AUDIO
SPEAKER A
CLEAN AUDIO
SPEAKER B
INTERFERENCE AUDIO
+
TRAINING TRIPLET
REFERENCE AUDIO
CLEAN AUDIO
NOISY AUDIO

TRAINING

THE DATA REQUIREMENTS FOR SPEAKER ENCODER AND VOICEFILTER ARE DIFFERENT SO THEY HAVE BEEN TRAINED SEPARATELY

SPEAKER ENCODER

- PUBLIC DATASETS (LibriSpeech, VoxCeleb)
- 34M UTTERANCES FROM 138K SPEAKERS

LibriSpeech is a corpus of approximately 1000 hours of 16kHz read English speech.

VoxCeleb : VoxCeleb is an audio-visual dataset consisting of short clips of human speech, extracted from interview videos uploaded to YouTube

VOICEFILTER

- PUBLIC TRANSCRIPED DATASETS (LibriSpeech, VCTK)
- DIVIDE EACH DATASET TO TRAINING AND EVALUATION SUBSETS

LibriSpeech is a corpus of approximately 1000 hours of 16kHz read English speech.

VCTK: Corpus includes speech data uttered by 110 English speakers with various accents. Each speaker reads out about 400 sentences.

EXPERIMENTAL RESULTS

WORD ERROR RATE (WER)

VOICEFILTER MODEL	CLEAN WER%	NOISY WER%
NO VOICE FILTER	6.1	60.6
VF TRAINED ON VCTK	21.1	37.0
VF TRAINED ON LibriSpeech	5.9	34.3

SOURCE TO DISTROTION RATIO (SDR)

VOICEFILTER MODEL	MEAN SDR (dB)	MEDIAN SDR (dB)
NO VOICE FILTER	10.1	2.5
USING VOICEFILTER	17.9	12.6

CONCLUSION

- WE PROPOSED A SPEAKER-CONDITIONED VOICE SEPARATION FRAMEWORK CALLED THE VOICE FILTER
- DEMONSTRATED THAT OUR SYSTEM HAS SIGNIFICANT WER IMPROVEMENT FOR MULTI-SPEAKER SCENARIOS AND MINIMAL DEGRADATION IN SINGLE-SPEAKER SCENARIOS
- THE PERFORMANCE CAN BE FURTHER IMPROVED BY USING MORE DATA.

THANK YOU