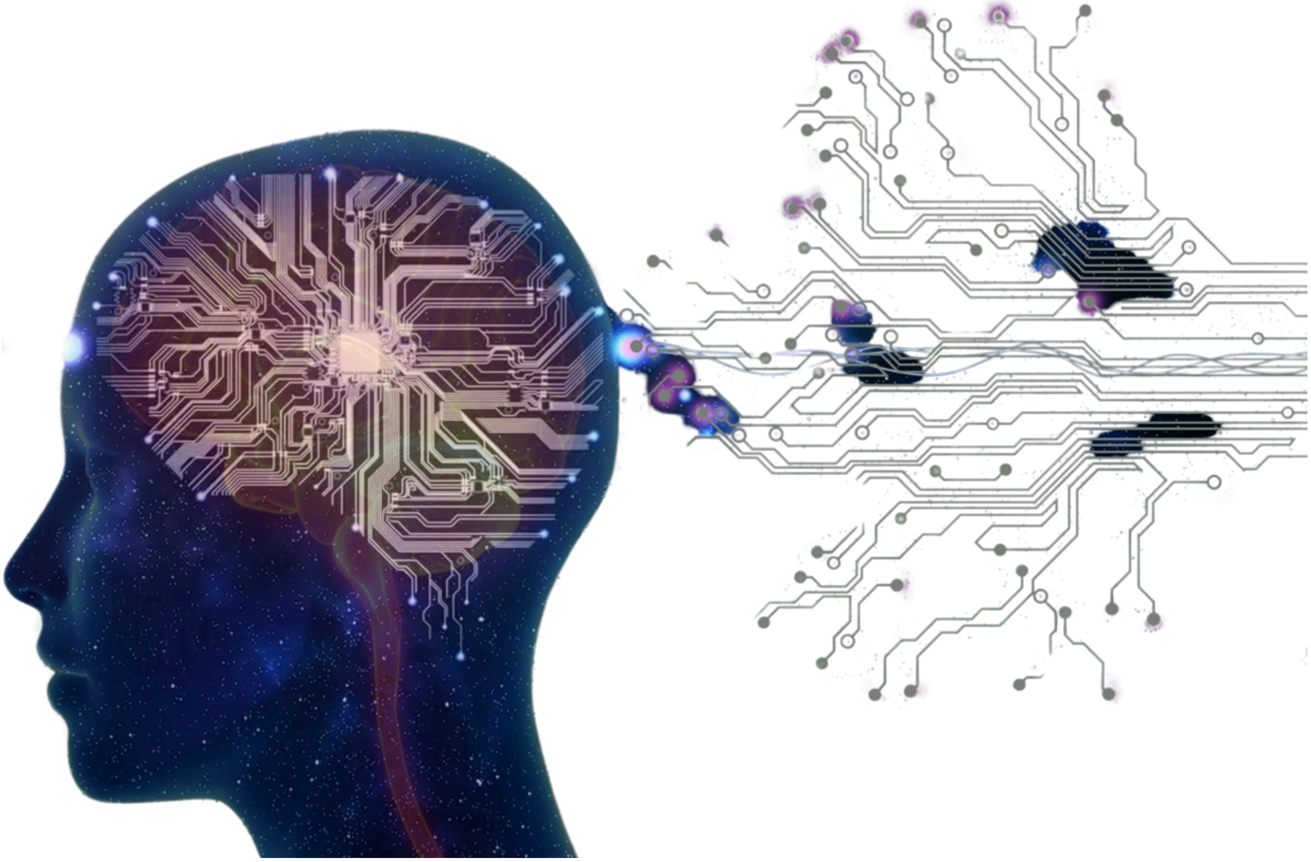
Meeami
TECHNOLOGIES

One Giant Leap for Audio Signal Processing

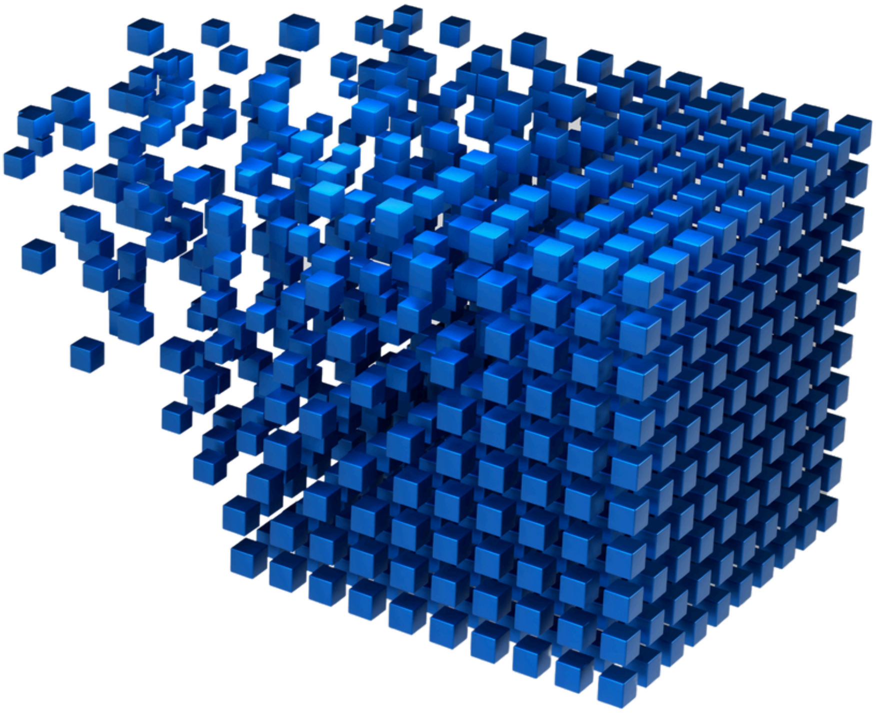


Contents

Contents

Part 1

Executive Summary



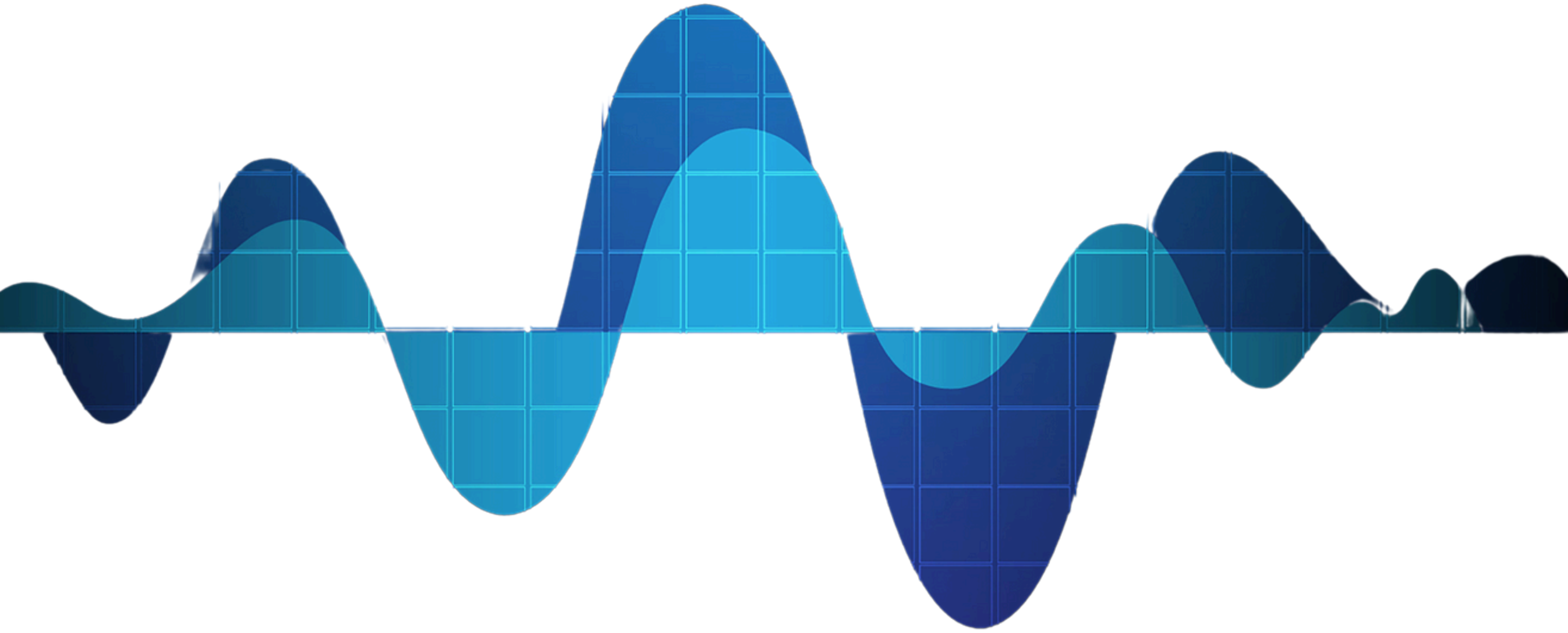
Introduction

This whitepaper aims to shine a light on how far audio signal processing has come since the advent of AI. The solutions have become more robust, have better performance, and when productized show lowered latency. This paper examines four of the most popular problem statements in audio signal processing: Noise Removal, Echo Removal, Packet Loss Concealment, and Speaker Identification. Their traditional solutions and AI-based solutions are given here. This paper also grazes the surface of some of the novel problems that AI has resolved. The applications of these problem statements are also outlined.

Meeami
TECHNOLOGIES

Part 2

Traditional Stochastic Signal Processing



Noise Removal

Any undesired signal can be considered as noise. It can degrade and subdue speech signals. Thus, it is imperative to remove noise to enhance speech.

Conventionally Wiener Filter and Kalman Filter have been used in conjunction (generally in 2 stages) with each other to remove noises from speech. In other instances, these filters can be used in conjunction with other adaptive filters (LMS, RMS filters, etc) to remove noise effectively.

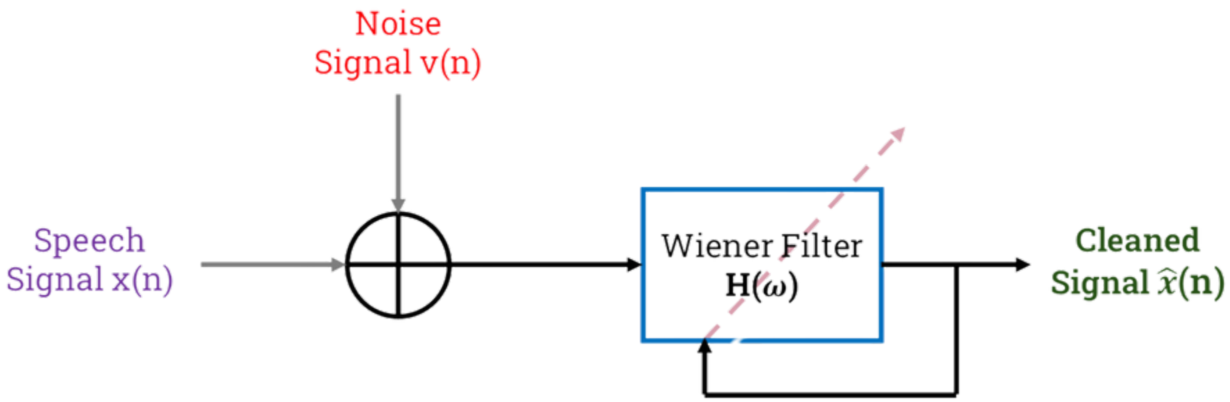
All noise removal occurs in the frequency domain.

Wiener Filter

A speech signal that has been corrupted with an Additive White Gaussian Noise (AWGN) can be cleaned by using a Wiener filter. Wiener filter works on the principle of optimising minimum mean square error between 2 random processes. It is traditionally a non-recursive filter. However, it is very popular as an adaptive filter and this adaptiveness has been illustrated here.

Underlying assumption: Original speech signal and AWGN are stationary linear processes with zero mean whose correlations (cross and auto) and power spectral densities are known.

The block diagram is as follows:



The minimized error is given as: $E\{|e\,(n)|^2\}$

where the error is defined as: $e\,(n) \approx \dot{x}\,(n) - x\,(n)$

The filter coefficients are found using the Wiener-Hopf equations in matrix form: $w = R^{-1}p$ where R^{-1} is the inverse Toeplitz matrix.

Pros: Works well for noises whose statistical properties change very little over time.
Ex: Fan noise, Steady hums of ACs and Refrigerators, Birds chirping, etc.

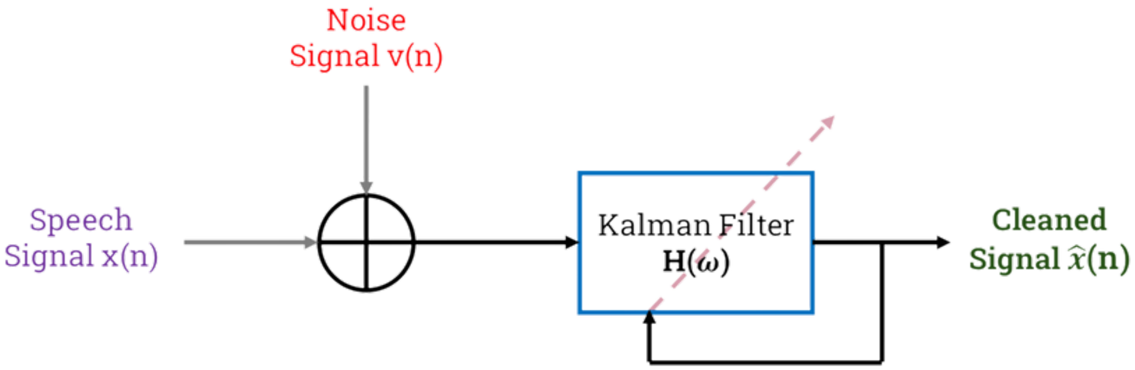
Drawbacks: Spectral properties of noises are not always known. They can also be difficult to calculate in the case of non-stationary and non-gaussian noises. Thus, it does not work well with dynamic noises.

Kalman Filter

A speech signal that has been corrupted with AWGN can be cleaned up by using a Kalman filter. Kalman filter works on the principle of optimising minimum mean square error between 2 random processes. Traditional Kalman filter is MMSE, Median, and MAP estimator all rolled into one single filter.[1]

Underlying assumption: Original speech signal and AWGN can be stationary/non-stationary linear processes whose correlations (cross and auto) and power spectral densities must be known.

Kalman filter works on minimizing error on the x(n) signal.



The filter coefficients are found out using the state space and time update equations with Kalman gain.

Pros: Works well for noises whose statistical properties are Gaussian. It is computationally faster than the Wiener filter.

Drawbacks: Spectral properties of noises are not always known. They can also be difficult to calculate in the case of non-stationary and non-gaussian noises. The initial state at n=0 might not be always known. Thus, it does not work well with dynamic noises.

Echo Removal

Acoustic echo pertains to the far-end signals that are captured along with near-end signals in the microphone of a device. The microphone and speaker are very near to each other and thus this phenomenon occurs. Echo and signal are coupled together in the microphone and thus it is a task to decouple them.

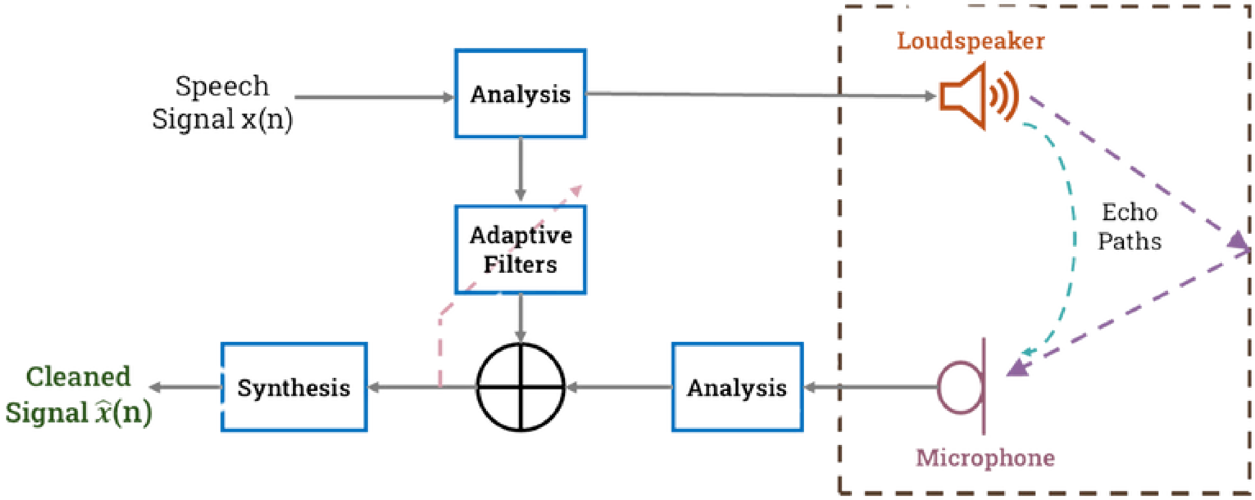
Conventional Echo Removal was achieved with LMS Algorithm and sub-band processing in conjunction with each other.

Least Mean Squares (LMS) Algorithm

Recursive normalized LMS is the algorithm that is implemented on an adaptive filter. The synthesis block provides the final cleaned speech signal.

The main parameters that are modeled are:

1. Room Impulse Response (RIR) - The impulse response here includes modeling the many pathways of echo and its delays.
2. Impulse Response of Speaker & Microphone
3. Setting the values of μ and $w'_n s$ for adaptive filters



The error is defined as: $e(n) \approx |y(n) - d(n)|$

The adaptive filter recursively estimates this error and thus echo is removed. The measure of echo removal is given by Echo Return Loss (ERL) in dB.

$$ERL = \frac{SignalPower}{EchoPower}$$

Negative or low ERL indicates that the near-end signal is buried in the far-end signal i.e. speech is buried in echo.

Higher ERL indicates good signal strength and lower echo.

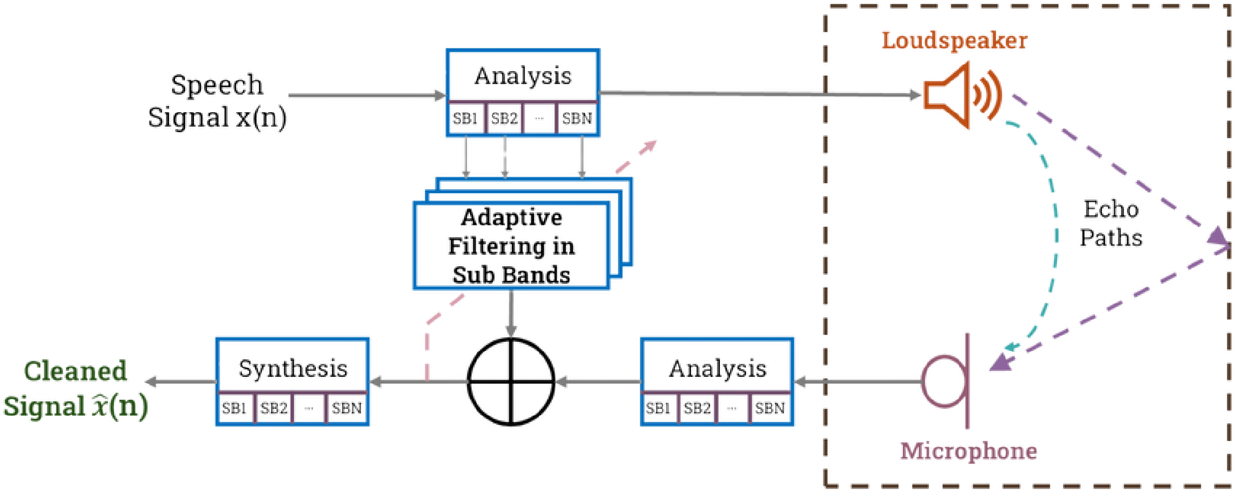
Pros: It is straightforward to implement and has low complexity.

Drawbacks: If the spatial position of the device speaker is changed in real-time, then the algorithm takes time to adjust to the new echo paths and the subsequent delays. Negative ERLs are not removed here.

Sub-Band Processing

In this approach, the signal is split into small time domain windows (Hann, Hamming, etc) and Short Time Fourier Transform is used to transform the signal into the Frequency domain. Adaptive filtering is applied separately in every frequency band to remove echo.

ERL is calculated as a measure of the strength of speech signal.



Pros: It is computationally faster (eigenvalue spread is lower - dynamically changing values are lesser) [5] and lighter than traditional LMS filtering since it does not process the entire sequence at once.

Drawbacks: Negative ERLs are not completely canceled here.

Packet Loss Concealment

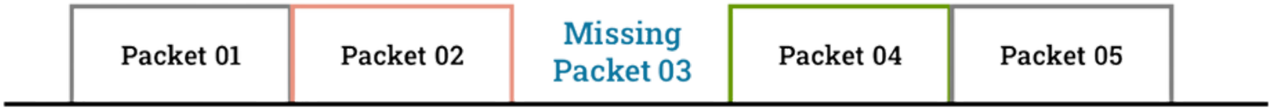
PLC occurs due to poor network conditions and poor hardware quality. Both can result in the loss of packets of speech. The loss of speech packets is detected by the receiver in the RTP header.

The conventional DSP method could conceal ~30-40 msec disparity at the maximum. If every packet is 20 msec, then max 2 packets could be concealed. If every packet was 10 msecs, then max 4 packets could be concealed.

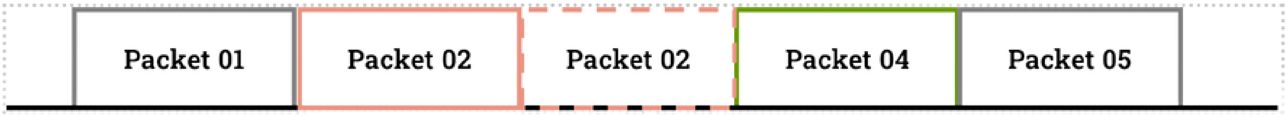
Waveform Similarity and Overlap Add (WSOLA) was the conventional method used. In this method, there were 3 ways to conceal the packet loss:

1. Unilateral Left WSOLA
2. Unilateral Right WSOLA
3. Bilateral WSOLA

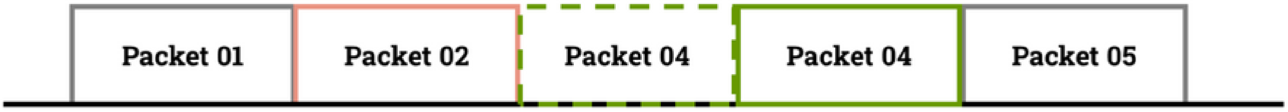
The original signal with packet loss is represented as follows:



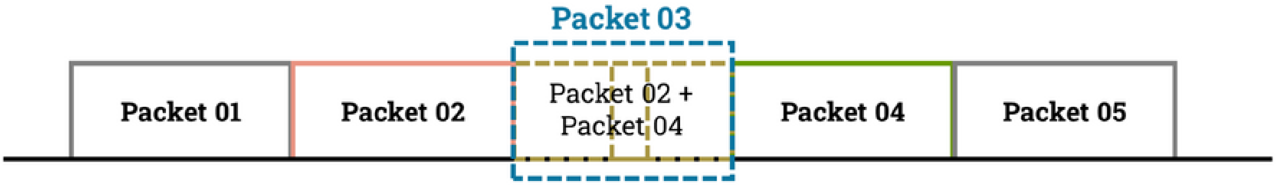
In unilateral left WSOLA, the leftmost packet to the lost packet is stretched (overlapped and added) and placed in the position of the lost packet. The other intact packets are placed after this. Another method of achieving this would be to stretch packet 01 & packet 02 to place packet 02 in place of packet 03. The first method is shown below:



In unilateral right WSOLA, the rightmost packet to the lost packet is stretched and placed in the position of the lost packet. The other intact packets are placed after this. Another method of achieving this would be to stretch packet 04 & packet 05 to place packet 04 in place of packet 03. The first method is shown below:



In bilateral WSOLA, the leftmost and rightmost packets to the lost packet are stretched and averaged. The resultant waveform is inserted in the place of the lost packet.



Pros: It is not an unpleasant experience for the human ear.

Drawbacks: It is not possible to compensate when too many consecutive packets are lost. This also results in loss of context in speech.

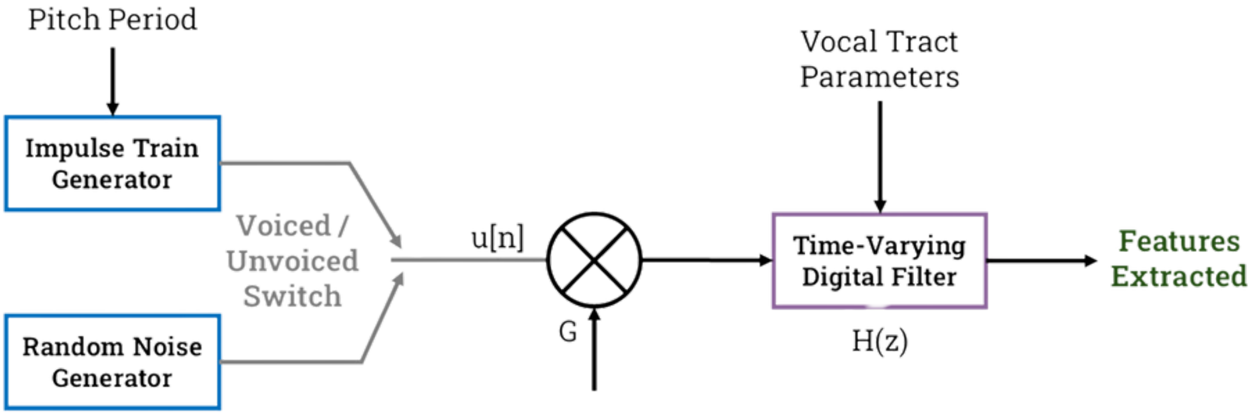
Speaker ID

The main goal of Speaker ID was to establish the identity of a human via voice. It could be dependent or independent of keywords. The principle of working was to extract the features of the speaker and classify them.

Conventionally LPC, MFCC's and HMM's were used in conjunction with each other.

Linear Predictive Coding (LPC)

LPC is used to extract features of speech. The current speech sample is approximated as a linear combination of past samples. The system is excited by an impulse train for voiced speech, or a random AWGN sequence for unvoiced speech. Windowing functions are used and a digital filter using minimum mean square error is used to predict the coefficients. This is a “source-filter” approach.

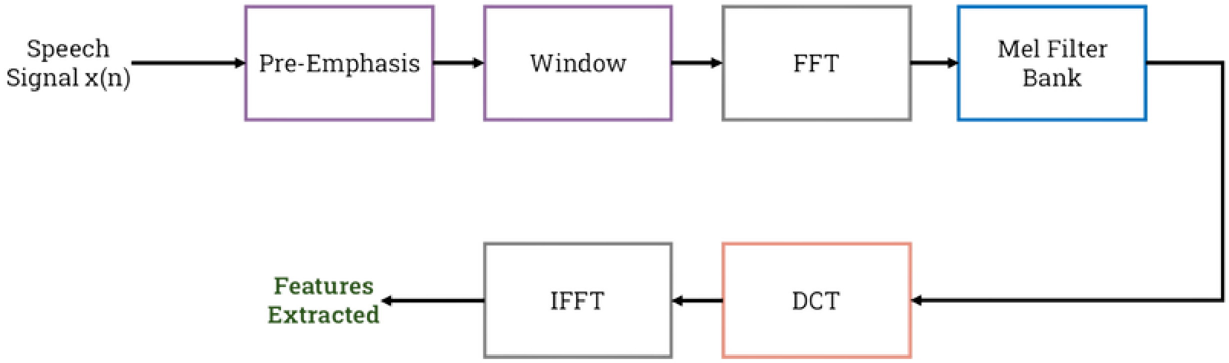


Pros: It is accurate for predicting the model of the linear system (vocal tract, glottal pulse, and radiation characteristic for voiced speech). It provides good-quality encoding at a low bit rate.

Drawbacks: All pole filters are not consistent with human speech production mechanisms or human hearing [4]. It is not very robust when speech is mixed with other background voices, echoes, and noises. It is not very robust when speech is mixed with other background voices, echoes, and noises.

Mel Frequency Cepstral Coefficients (MFCC)

MFCC is used to extract features of speech. Mel Frequency Cepstrum is the short-term power spectrum, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency. There are a series of digital overlapping filters that convert the signal into mel scale. These sets of filters are known as Mel filter banks. Mel-scale is linear till 1 KHz and logarithmic above it[3]. This closely resembles human ear perception when compared to other frequency scales. This is a “hearing-filter” approach.



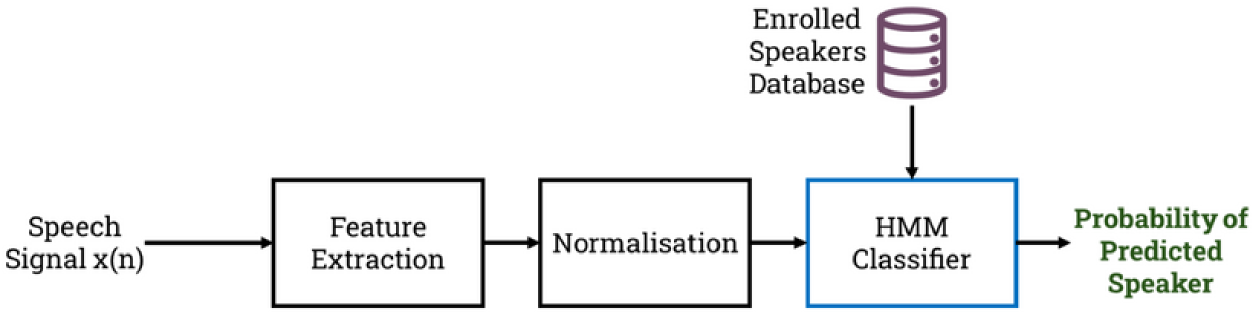
Pre-emphasis boosts the amount of energy in the high frequencies [2]. MFCC features that are extracted are generally independent of each other.

Pros: It is more robust than LPC.

Drawbacks: Performance is not adequate when speech is mixed with other background voices, echoes, and noises. It is also not very robust when the sounds are similar to each other.

Hidden Markov Models (HMM)

This is used as a classifier. LPC/MFCC extracts the features and HMM classifies them into classes. During the training phase, the features of known speakers are trained and classified according to HMM models. They are stored in a database. During implementation, the classifier compares data from the input to the data present in the database and makes a decision.



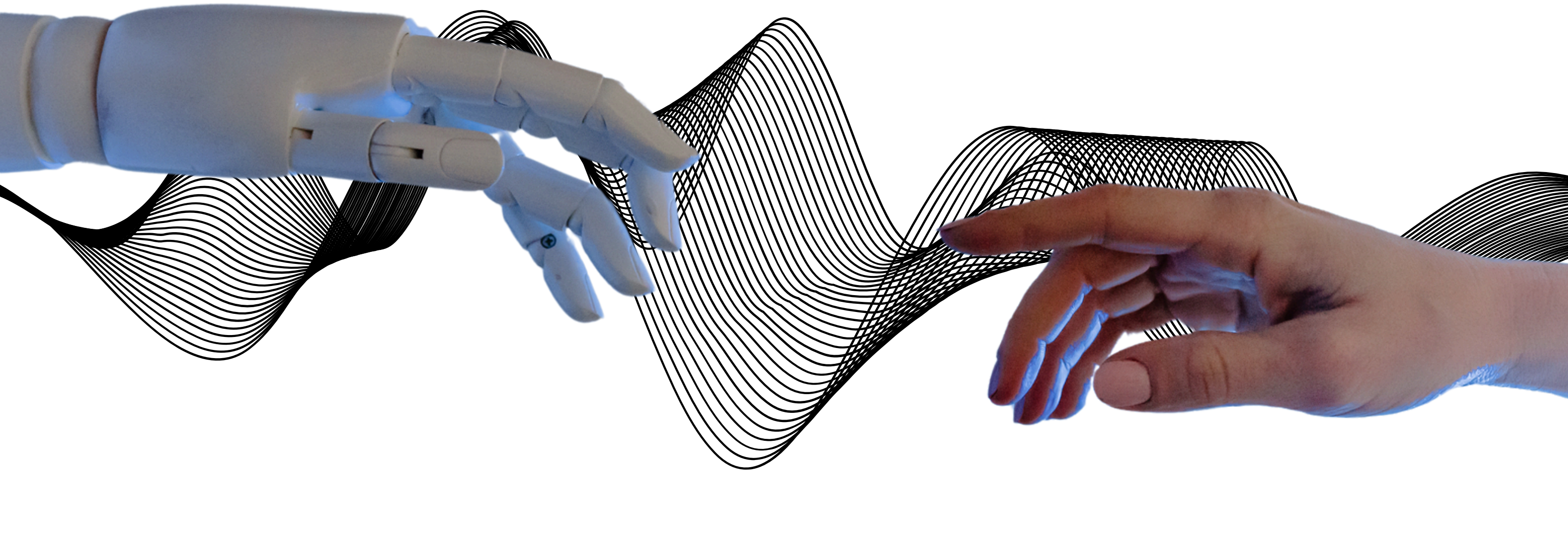
Pros: It provides a clear framework to interpret the model's behaviour and decisions.

Drawbacks: The features might not always be independent of each other. The features also might not always have a Gaussian distribution.

Meeami
TECHNOLOGIES

Part 3

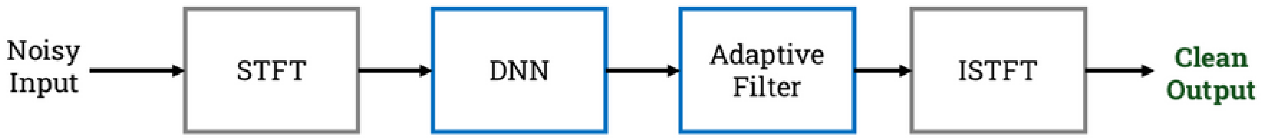
Improvements using AI



This section illustrates how the advent of AI has helped in the improvement of solutions of the above problem statements.

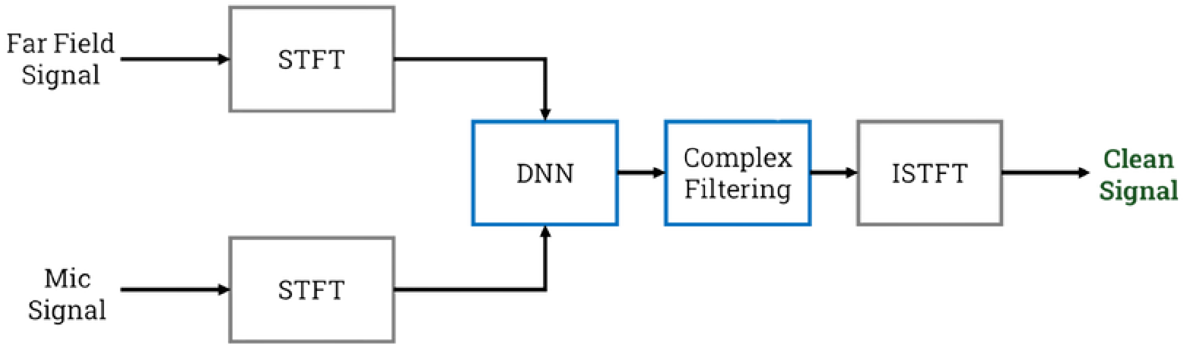
Noise Removal

The strict criterion of knowing the statistical properties of noises has been eliminated to a great extent. Noises and signals which do not follow Gaussian distribution can also be removed. Multitudes of dynamic noises can now be removed simultaneously in real time thus ensuring greater clarity of speech.



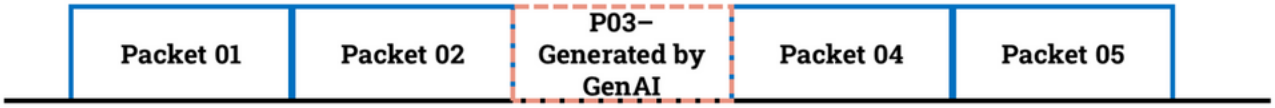
Echo Removal

The time lag required to adapt when the spatial position of the speaker has changed has lessened. Handling of low and negative ERLs has considerably improved.



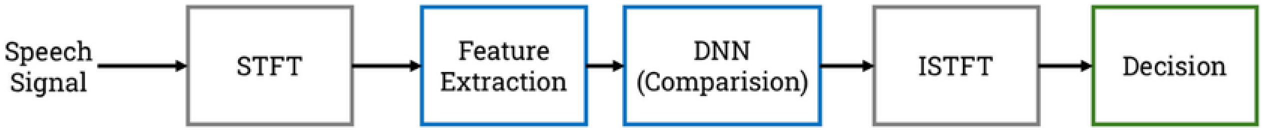
Packet Loss Concealment

Context and meanings of sentences are retained due to Recurrent Neural Networks (RNNs). Packets (Speech - Words) are generated using GenAI and inserted in the place of lost packets. If voice and accent changes are implemented on the generated speech, then the generated packet can be virtually indistinguishable from the lost packet.



Speaker ID

The dependency of features to have a Gaussian distribution has been eliminated. Unique identification is achievable despite similar features. It is possible to identify a human speaker with a greater degree of accuracy even in the presence of noises, echoes, and background voices.



Novel Problems Solved by AI

The following problem statements did not have solutions in traditional SSP. However, due to neural networks, it has been possible to design solutions.

Voice Suppression

It is now possible to teach the system to focus and enhance the primary speaker’s voice and suppress background voices irrespective of the amplitude levels of the background speakers.

Accent Change

It is now possible to extract features at a phoneme level. This enables modelling of various dialects and accents.

SD2HD

To change a Standard Definition (8 KHz / 16 KHz) signal (due to hardware limitations) to a Full Band High Definition (48 KHz) signal. This ensures greater speech intelligibility.

ASR and TTS

It is now possible to generate text from spoken speech (ASR). It is also possible to generate spoken speech (albeit robotic speech in raw form) from written text (TTS). It is also possible to do both in a plethora of vernaculars.

Augmentation of Datasets

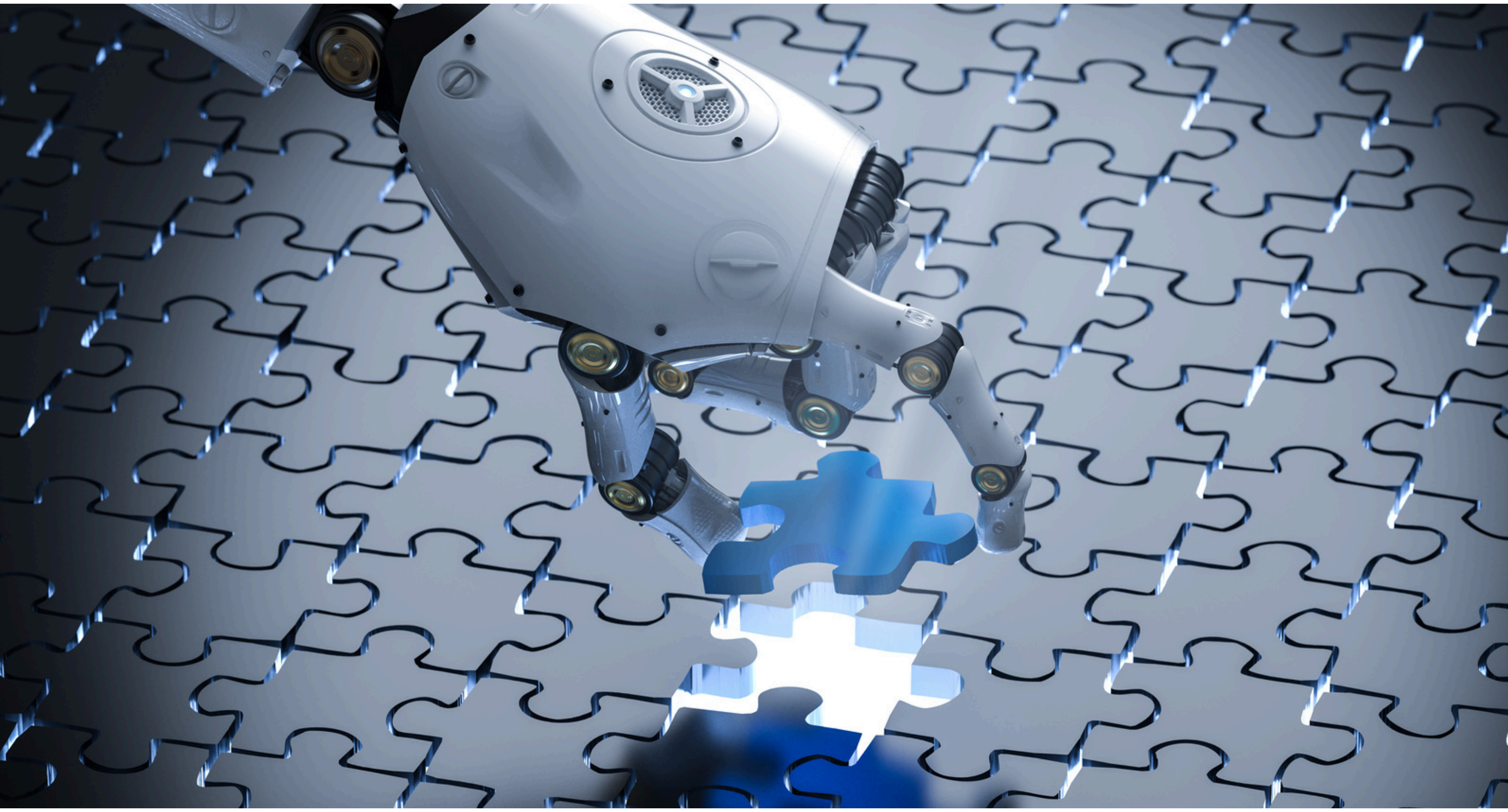
AI-based products have 2 phases - Training and Inference. It is imperative to have a good dataset for training. Augmentation of data to build datasets with numerous permutations and combinations factored in has led to better performance and accuracy during inference.

Ex: For background voice cancellation, it is merely enough to augment data instead of physically asking multiple people to speak in the mic at the same time.

Subsequently, this has also ensured that it is cost-effective to build powerful datasets.

Part 4

AI Applications in Audio



Applications

Some of the applications of the above solutions are given below. This is by no means an exhaustive list.

S.No	IP Name	Offline Solutions	Online Solutions
1	Noise Removal	Audio refinement in videos	Regular audio phone calls /customer support calls
		Historical Video Preservation	Video Calls - Live conferencing / gaming calls
2	Echo Removal	Audio refinement in videos	Regular audio phone calls /customer support calls
			Telemedicine
3	Packet Loss Concealment	Audio Forensics	Emergency Calls
		Historical Video Preservation	Telemedicine
4	Speaker ID	Home/Office/School Speaker ID	Banking Calls
		Named Transcriptions	Live Transcriptions
5	Accent Change	Accent change of instructor in recorded classes	Customer support calls
		Personalized Audiobooks and Narration	Webinars / Conference Calls
6	SD2HD	Historical Video Preservation	Live Conference calls
7	ASR	General Transcriptions	Patient's / Doctor's Notes & Paperwork
8	TTS	Text to speech readers	Automated Chat Bots

Emerging Possibilities

As illustrated above, non-AI-based traditional SSP products have drawbacks related to performance. The advent of AI has thus enabled signal processing to take a giant leap in terms of increasing performance for existing solutions and providing novel solutions to hitherto unsolved problems.

As of the date on which this article is being written, the following problem statements still do not have widespread commercially cost-effective solutions:

1. Cross-Language Speech Synthesis
2. Selective enhancement (Event Detection/Voice) amidst ambient background (Noise + Voices + Echoes)

AI research into these domains can prove to be pathbreaking in terms of breaking communication barriers and audio forensics.

References

[1] https://dsp.stackexchange.com/questions/66222/kalman-filter-difference-between-minimizing-the-mean-square-error-mmse-max
[2] https://jonathan-hui.medium.com/speech-recognition-feature-extraction-mfcc-plp-5455f5a69dd9
[3] https://en.wikipedia.org/wiki/Mel-frequency_cepstrum
[4]https://www.researchgate.net/publication/261914482_Feature_extraction_methods_LPC_PLP_and_MFCC_in_speech_recognition
[5] https://vocal.com/echo-cancellation/sub-band-acoustic-echo-cancellation/