A Simple Overview of Monaural Speech Enhancement

——focusing on the case of additive independent noise

ChenChen

2021.5.20

目录

- 问题描述
- 方法分类
- 典型算法

问题描述

- 背景
  - 语音混杂着噪声——如何获取干净的语音？
- 目的
  - perceived quality 感知质量
  - intelligibility 清晰度和可理解性
- Our focus
  - denoise

数学模型

- 前提
  - the noise is additive and independent of the clean speech
- 模型
  - 带噪语音信号序列是原始语音信号序列和噪音信号序列之和

$$s(k)=f\left(s(k-1),\cdots,s(k-K),\mathbf{w}\right)+v(k)$$
$$y(k)=s(k)+n(k)$$

语音的特性

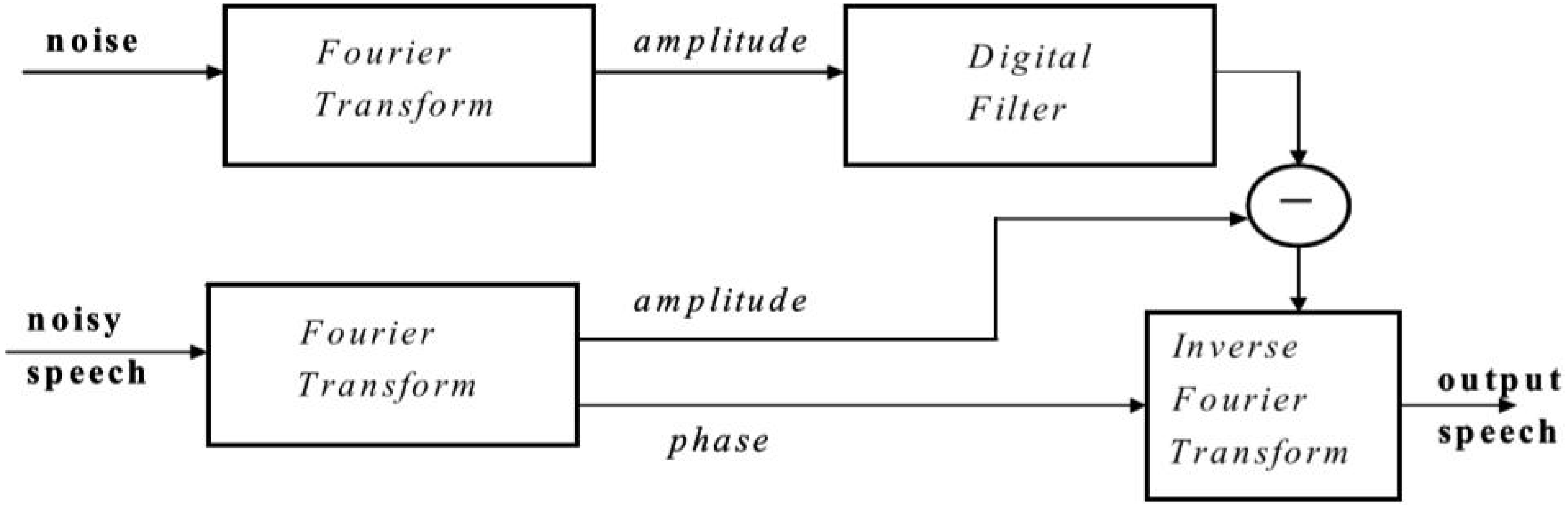
- 非线性&非平稳性
  - 音素内和音素之间的快速片段过渡
  - 清音期间的湍流激励
  - 浊音期间的声门打开/闭合
- 线性&平稳性
  - 短时线性时不变
  - 短时平稳信号

方法分类

- Spectral Subtraction/Filtering Techniques
  - Spectral Subtraction (SS)
  - Wiener filtering
  - Kalman Filtering
  - Signal Subspace approach
- Neural Network Based Techniques

Spectral Subtraction (SS)

- 前提：Speech and noise are assumed to be uncorrelated
- 思想：整体的短时能量谱减去噪声的短时能量谱



Spectral Subtraction (SS)

- 优势
  - 简单、有效、直观

$$\left|\hat{S}(\omega)\right|^2 = \begin{cases} \left|Y(\omega)\right|^2 - \left|\hat{N}(\omega)\right|^2 & if \quad \left|Y(\omega)\right|^2 > \left|\hat{N}(\omega)\right|^2 \\ 0 & otherwise \end{cases}$$

$$\hat{s}(k) = IFFT\left[\left|\hat{S}(\omega)\right| e^{j\arg(Y(\omega))}\right]$$

- 缺陷
  - 没有估计原始相位信息，而使用混合信号的相位信息做逆变换，造成失真
  - 对相减结果出现的负值直接置零，转换回时域后引入“音乐噪声”
- 拓展方向
  - Spectral Subtraction With Oversubtraction Model
  - Non-Linear Spectral Subtraction

Spectral Subtraction (SS)

- Spectral Subtraction With Oversubtraction Model
  - α :过减因子
  - β :谱下限

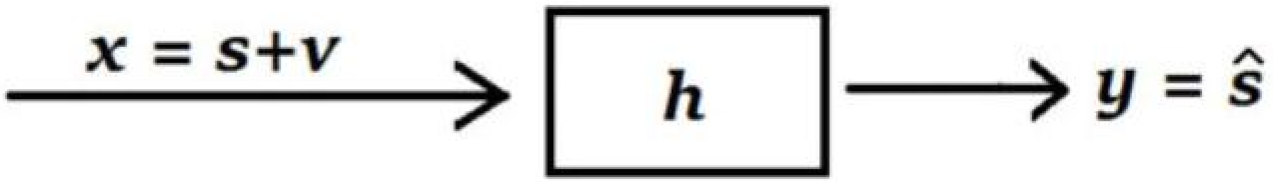
$$|\hat{S}(\omega)|^2 = \begin{cases} |Y(\omega)|^2 - \alpha|\hat{N}(\omega)|^2 \text{ if } |Y(\omega)|^2 - |\hat{N}(\omega)|^2 > \beta|\hat{N}(\omega)|^2 \\ \beta|\hat{N}(\omega)|^2 \text{ otherwise} \end{cases}$$

- Non-Linear Spectral Subtraction
  - Φ is a non-liner function

$$\left|\hat{N}(\omega)\right|_{nl}^2 = \Phi\left(\max_{over\ M\ frames}\left(\left|\hat{N}(\omega)\right|^2\right), R_{post}(\omega), \left|\hat{N}(\omega)\right|^2\right)$$

$$R_{post}(\omega) = \left(\left|Y(\omega)^2\right| / \left|N(\omega)^2\right|\right) - 1$$

Wiener filtering



- 前提：
  - 语音和噪声均为广义平稳过程且知它们的二阶统计特性
- 思想：
  - 利用信号和噪音的自相关函数来获得最小均方误差(MMSE)意义下对线性滤波器最优预测
    - 本质上是一个线性最小均方差估计器（LMMSE estimator）
- 限制
  - 维纳滤波器是在一维平稳状态下的线性最优估计器，只有输入信号是统计意义上是平稳信号时，其增益函数解才是最优解
  - 仅仅考虑了量测方程，并没有关心信号本身的变化规律

Kalman filtering

- 前提
  - 动力学模型是线性的，量测模型也是线性的
  - 状态噪声和量测噪声均为零均值的白噪声
  - 两种噪声，以及噪声与状态之间互不相关
- 思想
  - 粗略地讲，Kalman filter就是一种可以recursively执行的，结合了线性系统动态方程的Wiener filter。

Extended Kalman filtering

$$s(k)=f\left(s(k-1),\cdots,s(k-K),\mathbf{w}\right)+v(k)$$
$$y(k)=s(k)+n(k)$$

- 模型
  - 语音时域模型为非线性自回归模型
    - v(k) 是状态方程中的过程噪音，通常认为是白噪声
- 思想
  - 用一个时变线性函数作为非线性函数f(·)的近似

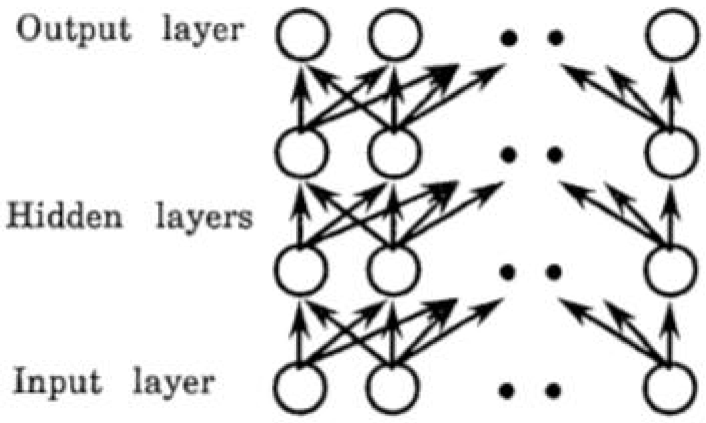
Signal Subspace approach

- 前提
  - Assuming the signal and noise are stationary
- 思想
  - Decompose the vector space of the noisy signal into a signal-plusnoise subspace and a noise subspace.
  - Enhancement is performed by removing the noise subspace and estimating the clean signal from the remaining signal subspace.
- 缺陷：
  - 子空间正交的假设在实际情况下并不精确
  - 对非平稳噪声的效果较差

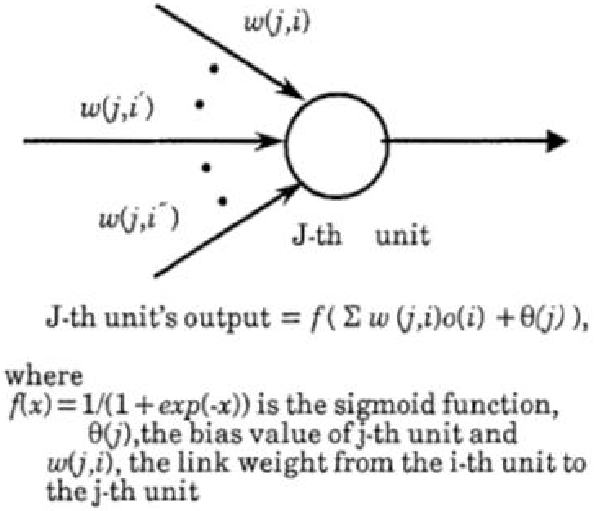
Neural Network Based Approaches

- Neural Networks as **nonlinear filters** mapping the noisy speech to clean speech in the time domain or in different domains
- A **time variant model** can be achieved by creating **different fixed models** for corresponding dynamical regimes of the signals and **switching between these models** during the speech enhancement process.

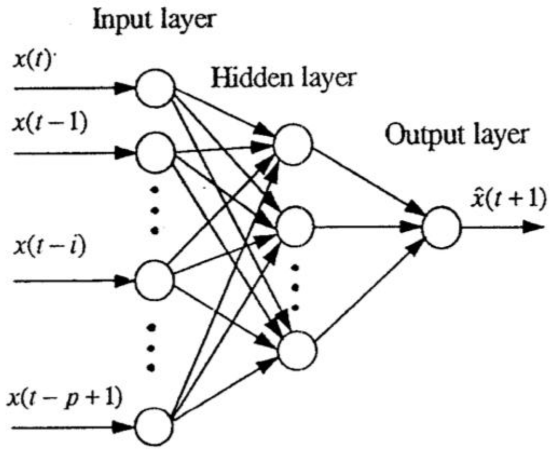
The Tamura approach



- 前提
  - availability of a clean speech training set
  - additive noise (non-stationary)
- 结构
  - the input and output of the network is given by **the waveform itself**, the units on the output and input layers are all **linear units**
  - **Learning by Error Back-Propagation**
- 缺陷
  - Attenuates many high frequency components in the actual speech



Speech Signal Restoration Using an Optimal Neural Network Structure



- 目标
  - Select the optimal coniplexity of the network structure so that the network can remove the noise without distorting the original speech signal
- 思路
  - Use a **feedforward neural network with one hidden layer** as a nonlinear predictive filter.
  - The hyperbolic tangent functions are used as the **nonlinear transfer function of the hidden nodes** and **the transfer function of the output layer node is linear.**
  - Apply the **Predictive Minimum Description Length (PMDL) principle** to determine the optimal number of input and hidden nodes.

NPHMM Neural Predictive Hidden Markov Model

- 前提
  - the nonlinear and nonstationarity nature of speech
- 思路
  - NPHMM is a **nonlinear autoregressive process** whose time-varying parameters are controlled by a **hidden markov chain**, speech is the output of a NPHMM.
  - Given some speech data for training, the parameter of NPHMM is estimated by a learning algorithm based on the combination of **Baum-Welch algorithm** and a neural network learning algorithm using the **back propagation algorithm**.
  - The Extendend Kalman Filter (EKF) technique, involving an autoregressive model for each class, can be used to provide the maximum-likelihood estimation for speech.

Denoise Auto Encoder

- 前提
  - DO NOT require any such apriori conditions to be met when applying the enhancement
- 结构
  - Use a deep neural network (DNN) with multiple layers of fully connected neurons
- 目标
  - Estimate the masks that give the desired clean speech spectra after multiplying the noisy spectra (masking)
  - Estimate clean speech spectra directly (mapping)
- 拓展
  - CDAE
    - the use of a convolutional neural network (CNN) as a convolutional denoising autoencoder

Other NN Based Approaches

- RNPHMM(Recurrent Neural Predictive Hidden Markov Model)
  - The nonlinear prediction model is based on a Recurrent Neural Network
  - The unknown parameters are estimated by a learning algorithm derived from the Baum-Welch and RNN back-propagation algorithms
- Employing time delay neural network for Mel-scaled spectral estimation
- Multi-layer perceptron (MLP) neural network estimate the log spectra of speech
- Dual EKF
  - A neural network based time-domain method removing nonstationary and colored noise from speech.
  - the availability of only the noisy signal