# Combined speech enhancement and auditory modelling for robust distributed speech recognition

Author : Ronan Flynn, Edward Jones

Professor: 陳嘉平

Reporter: 吳國豪

# Outline

- Introduction
- Front-end processing
- Speech enhancement
- Experiments

# Introduction

- It is well-known that the presence of noise severely degrades the performance of speech recognition systems.
  - One common approach to improving system performance in noise is to use front-ends that produce robust features.
  - Another method that has been proposed to improve the robustness of ASR systems is to enhance the speech signal before feature extraction.

# The auditory model of Li

- The auditory feature extraction algorithm proposed by Li et al. is based on an analysis of **the human auditory system**. The steps involved in the feature extraction are shown in Fig. 1.

- An **outer/middle ear transfer function** (see Fig. 2) that models pressure gain in the outer and middle ears is applied to the spectrum magnitude. The spectrum is then subjected to a non-linear frequency transformation to convert it to the **Bark scale**.

- After conversion of the spectrum to the Bark scale, the transfer function output is processed in the frequency domain by an **auditory filter** that is derived from **psychophysical measurements** of the frequency response of the **cochlea**.

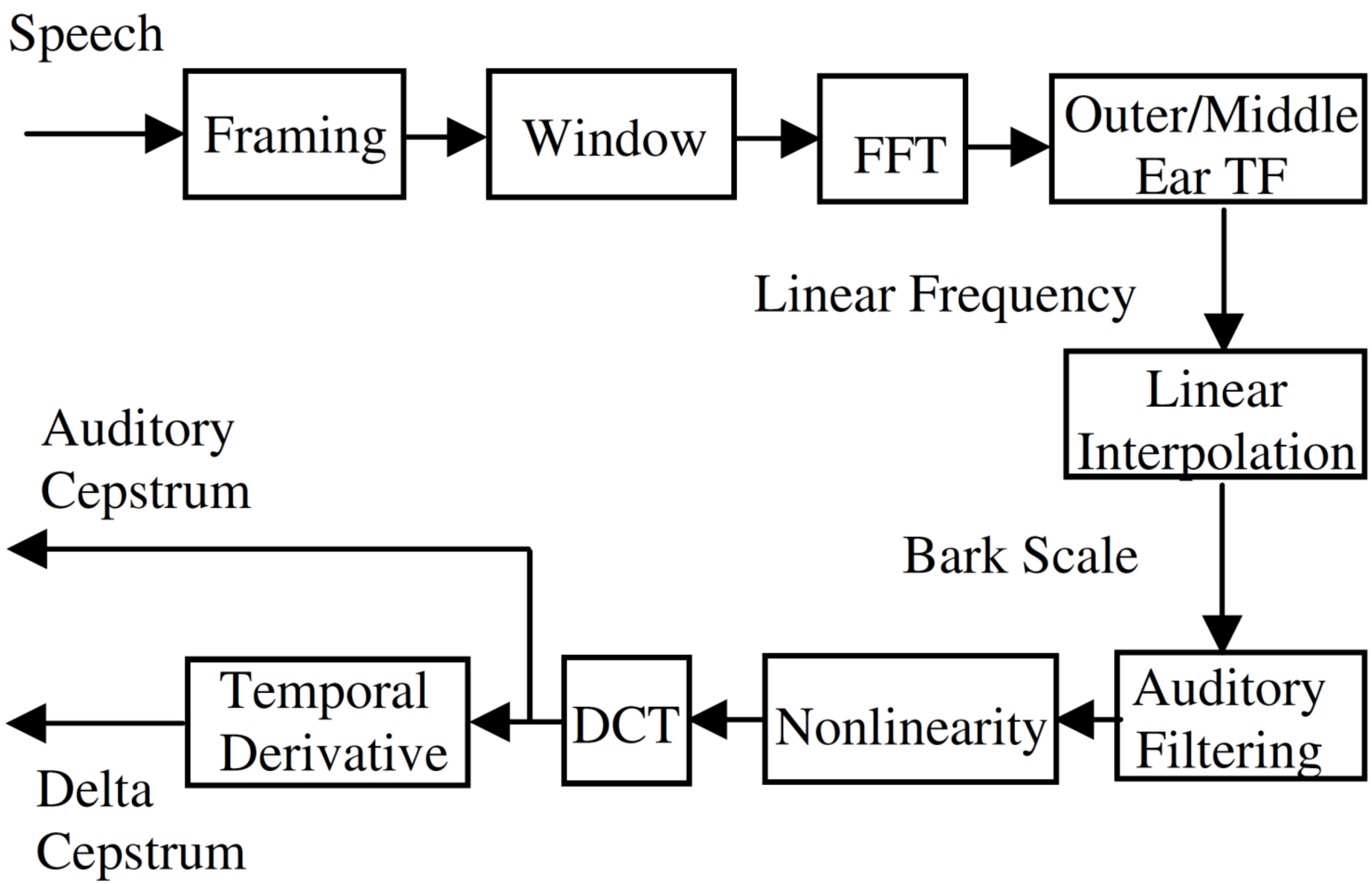


Fig. 1. Feature extraction proposed by Li et al.

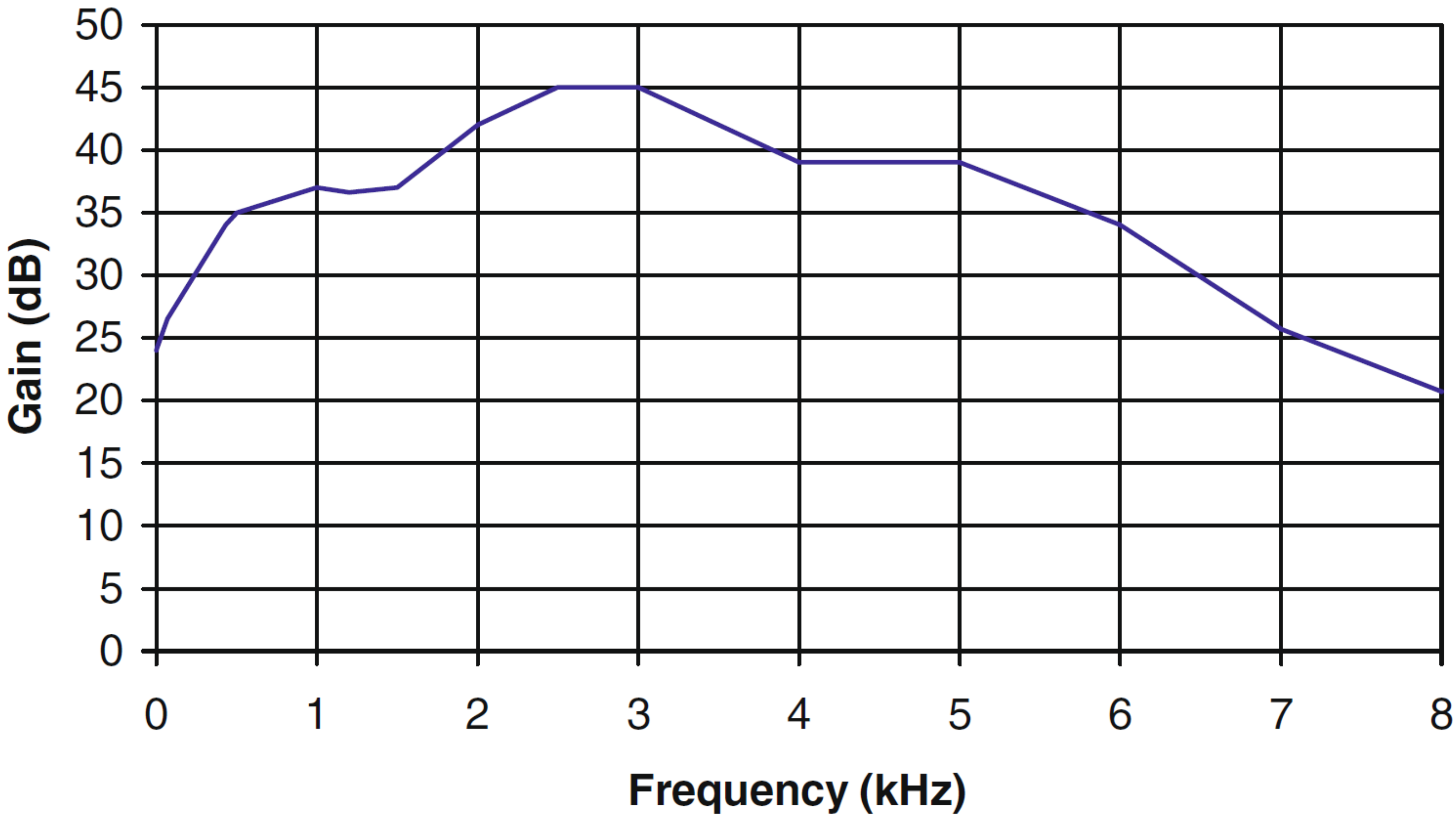


Fig. 2. Outer/middle ear transfer function

# The auditory model of Li

- Like the MFCCs, a non-linear function in the form of a **logarithm**, followed by a DCT, is applied to the filter outputs to generate the cepstral coefficients.

- The first generates a feature vector consisting of 13 coefficients made up of the frame log-energy measure and the cepstral coefficients $C_1$ to $C_{12}$.

- The second version generates a feature vector that contains the cepstral coefficients $C_1$ to $C_{12}$ along with a weighted combination of cepstral coefficient $C_0$ and the frame log-energy measure.

# Ephraim and Malah

- Ephraim and Malah present a minimum mean square error short-time spectral amplitude (MMSE STSA) estimator.
- In a noisy signal x(t), the MMSE amplitude estimator of the *k*th spectral component is given by

$$\hat{A}_k = G_k R_k$$

where $R_k$ is the amplitude of the *k*th spectral component in x(t) and $G_k$ is given by

$$G_k = \frac{\sqrt{\pi}}{2} \cdot \frac{\sqrt{v_k}}{SNR_{post\_k}} . M[v_k]$$

# Ephraim and Malah

- $v_k$ is calculated as $$v_k = (\frac{SNR_{prio\_k}}{1+SNR_{prio\_k}}).SNR_{post\_k}$$

  where $SNR_{prio\_k}$ and $SNR_{post\_k}$ are the a priori and a posteriori signal-to-noise ratios, respectively.

- The function M[ ] is evaluated as follows:

$$M[\theta] = \exp(\frac{-\theta}{2})[(1+\theta)I_0(\frac{\theta}{2}) + \theta I_1(\frac{\theta}{2})]$$

$I_0$ and $I_1$ represent the modified Bessel functions of zero and first-order, respectively.

# Ephraim and Malah

- The *a priori* SNR for the *k*th spectral component in the *n*th analysis frame is determined by

$$SNR_{prio\_k}(n) = \alpha\left(\frac{\hat{A}_k^2(n-1)}{\lambda_k(n-1)}\right) + (1-\alpha)\,p[SNR_{post\_k}(n)-1]$$

where $0 \le \alpha < 1$, $\lambda_k$ is the variance of the *k*th spectral component of the noise and P[ ] is a half-wave rectification operator which is defined by

$$p[x] = \begin{cases} x \ \ if \ \ x \ge 0, \\ 0 \ \ otherwise. \end{cases}$$

- The a posteriori signal-to-noise ratio, $SNR_{post\_k}(n)$ is determined using $(R_k)^2$, the amplitude-squared of the *k*th spectral component, and the current estimate of the noise power.

# Westerlund

- Westerlund et al. present a speech enhancement technique in which the input signal is first divided into **a number of sub-bands**. The signal in each sub-band is individually multiplied by a **gain factor** in the time-domain based on an estimate of the **short-term SNR** in each subband at every time instant.

- Westerlund et al. consider a discrete time speech signal, s(n), corrupted by a noise signal, w(n), that results in a noise corrupted speech signal x(n), where

$$x(n) = s(n) + w(n).$$

# Westerlund

- After filtering x(n) by a bank of k bandpass filters, x(n) can be written as

$$x(n)=\sum_{k=0}^{k-1}x_k(n)=\sum_{k=0}^{k-1}s_k(n)+w_k(n)$$

where $x_k(n)$ is the sub-band noisy speech signal. Westerlund et al. calculate a gain function, $g_k(n)$ , for each sub-band and this function weights the input signal subbands based on the ratio of $s_k(n)$ to $w_k(n)$ . The enhanced signal is given by

$$y(n)=\sum_{k=0}^{k-1}g_k(n)x_k(n)$$

# Westerlund

- In each sub-band, the short-term exponential magnitude average, $A_{x,k}(n)$ is based on $|x_k(n)|$ ; and an estimate of the noise floor level, $\underline{A}_{x,k}(n)$ , are calculated according to the following equations.

$$A_{x,k}(n) = (1-\alpha_k)A_{x,k}(n-1) + \alpha_k \left|x_k(n)\right|$$

$$\underline{A}_{x,k}(n) = \begin{cases} (1+\beta_k)\times \underline{A}_{x,k}(n-1) \ \ if \ \ A_{x,k}(n) > \underline{A}_{x,k}(n-1) \\ A_{x,k}(n) \ otherwise \end{cases}$$

$$g_k(n) = \left(\frac{A_{x,k}(n)}{\underline{A}_{x,k}(n)}\right)^{p_k}, \ p_k \geq 0, \ \underline{A}_{x,k}(n) > 0$$

$$g_k(n) = \begin{cases} g_k(n) \ if \ \ g_k(n) \leq L_{k,} \\ L_k \ \ otherwise. \end{cases}$$

# Rangachari and Loizou

- The smoothed power spectrum of the noisy speech signal is given by

$$P(\lambda,k)=\eta P(\lambda-1,k)+(1-\eta)\left|y(\lambda,k)\right|^2$$

where $\lambda$ is the frame index, k the frequency index, $\eta$ a smoothing constant and $|y(\lambda,k)|^2$ is the short-time power spectrum of the noisy speech.

- The local minimum of the noisy speech power spectrum, $p_{\min}(\lambda,k)$ is give by

$$P_{\min}(\lambda,k)=\begin{cases}\gamma P_{\min}(\lambda-1,k)+\dfrac{1-\gamma}{1-\beta}(P(\lambda,k)-\beta P(\lambda-1,k))\ if\ P_{\min}(\lambda-1,k)<P(\lambda,k)\\ P(\lambda,k)\ otherwise\end{cases}$$

# Rangachari and Loizou

- The ratio of the noisy speech power spectrum to its local minimum:

$$if \quad \frac{P(\lambda,k)}{P_{\min}(\lambda,k)} > \delta(k) \; I(\lambda,k) = 1 \;\; speech \; present$$
$$else \; I(\lambda,k) = 0 \; speech \; absent$$

- The speech-presence probability is updated as follows:

$$p(\lambda,k) = \alpha_p p(\lambda-1,k) + (1-\alpha_p) I(\lambda,k)$$

# Rangachari and Loizou

- The noise power spectrum estimate, $D(\lambda,k)$ , is then updated as

$$D(\lambda,k) = \alpha_s(\lambda,k)D(\lambda-1,k) + (1-\alpha_s(\lambda,k))\left|y(\lambda,k)\right|^2$$

$$\alpha_s(\lambda,k) = \alpha_d + (1-\alpha_d)\,p(\lambda,k)$$

- The estimated clean speech spectrum is evaluated as

$$C(\lambda,k) = \max\{\left|y(\lambda,k)\right|^2 - D(\lambda,k), vD(\lambda,k)\}$$

# Experiments

- Aurora 2

- The Aurora database also contains noisy data. This corresponds to clean data with noise artificially added at SNRs of 20 dB, 15 dB, 10 dB, 5 dB, 0 dB and 5 dB.

Recognition results – Li et al. (I)

| Enhancement | Absolute word accuracy % | | | |
|---|---|---|---|---|
| | Set A | Set B | Set C | Overall |
| None | 62.16 | 64.31 | 57.76 | 62.14 |
| Ephraim and Malah | 78.85 | 79.38 | 74.78 | 78.25 |
| Westerlund et al. | 75.87 | 76.32 | 70.45 | 74.97 |
| Rangachari and Loizou | 74.50 | 73.16 | 74.29 | 73.92 |

Recognition results – ETSI basic front-end

| Enhancement | Absolute word accuracy % | | | |
|---|---|---|---|---|
| | Set A | Set B | Set C | Overall |
| None | 61.34 | 55.75 | 66.14 | 60.06 |
| Ephraim and Malah | 76.34 | 75.91 | 73.71 | 75.64 |
| Westerlund et al. | 76.04 | 72.54 | 72.36 | 73.90 |
| Rangachari and Loizou | 63.58 | 61.57 | 67.82 | 63.62 |

Recognition results – Li et al. (II)

| Enhancement | Absolute word accuracy % | | | |
|---|---|---|---|---|
| | Set A | Set B | Set C | Overall |
| None | 67.34 | 69.18 | 63.44 | 67.30 |
| Ephraim and Malah | 80.36 | 81.03 | 79.34 | 80.42 |
| Westerlund et al. | 78.70 | 80.02 | 78.44 | 79.18 |
| Rangachari and Loizou | 76.08 | 76.16 | 75.94 | 76.08 |

Recognition results – ETSI advanced front-end

| Enhancement | Absolute word accuracy % | | | |
|---|---|---|---|---|
| | Set A | Set B | Set C | Overall |
| None | 65.92 | 65.48 | 70.07 | 66.57 |
| Ephraim and Malah | 77.92 | 77.61 | 78.64 | 77.94 |
| Westerlund et al. | 79.09 | 79.13 | 79.70 | 79.23 |
| Rangachari and Loizou | 73.77 | 73.35 | 78.85 | 74.62 |