Robust Speech Recognition Using a Cepstral Minimum-Mean-Square-Error-Motivated Noise Suppressor

Author: Dong Yu, Li Deng, Jasha Droppo,
Jian Wu, Yifan Gong, Alex Acero

Professor:陳嘉平

Reporter:葉佳璋

Abstract—We present an efficient and effective nonlinear feature-domain noise suppression algorithm, motivated by the minimum-mean-square-error (MMSE) optimization criterion, for noise-robust speech recognition. Distinguishing from the log-MMSE spectral amplitude noise suppressor proposed by Ephraim and Malah (E&M), our new algorithm is aimed to minimize the error expressed explicitly for the Mel-frequency cepstra instead of discrete Fourier transform (DFT) spectra, and it operates on the Mel-frequency filter bank's output. As a consequence, the statistics used to estimate the suppression factor become vastly different from those used in the E&M log-MMSE suppressor. Our algorithm is significantly more efficient than the E&M's log-MMSE suppressor since the number of the channels in the Mel-frequency filter bank is much smaller (23 in our case) than the number of bins (256) in DFT. We have conducted extensive speech recognition experiments on the standard Aurora-3 task. The experimental results demonstrate a reduction of the recognition word error rate by 48% over the standard ICSLP02 baseline, 26% over the cepstral mean normalization baseline, and 13% over the popular E&M's log-MMSE noise suppressor. The experiments also show that our new algorithm performs slightly better than the ETSI advanced front end (AFE) on the well-matched and mid-mismatched settings, and has 8% and 10% fewer errors than our earlier SPLICE (stereo-based piecewise linear compensation for environments) system on these settings, respectively.

Introduction

- We proposed nonlinear feature-domain noise reduction algorithm motivated by the minimum-mean-square-error(MMSE) criterion on MFCC
- We derive the algorithm by
  - ➢Assigning uniformly distributed random phase to the real-valued filter bank's outputs
  - ➢Assuming that the artificially generated complex filter bank's outputs follow zero-mean complex normal distributions

Problem Formulation

- We assume that x(t) is a corrupted with independent additive noise waveform n(t) become the noisy speech waveform, i.e.

$$y(t) = x(t) + n(t)$$

- We get the relationship in the DFT domain

$$Y(f) = X(f) + N(f)$$

- The Mel-frequency filter bank's output power for noisy feature

$$m_y(b) = \sum_f \omega_b(f) \left| Y(f) \right|^2$$

- The kth dimension of MFCC is calculated as

$$c_y(k) \cong \sum_b a_{k,b} m_y(b) \qquad a_{k,b} = \cos \frac{\pi b}{B}(k - 0.5)$$

Problem Formulation

- Our goal is to find the MMSE estimate $\hat{c}_x(k)$ against to each separate and independent dimension k of MFCC vector c_x

$$\hat{c}_x(k) = \hat{f}\left(c_y(k)\right) = \arg\min_f E\left\{\left(f\left(c_y(k)\right) - c_x(k)\right)^2\right\}$$

$$= \arg\min_f \int \left(f\left(c_y(k)\right) - c_x(k)\right)^2 p\left(c_x(k)\right) dc_x(k)$$

- Three reasons for choosing the dimension-wise instead of full-vector MMSE criterion
  - ➢ Each dimension of MFCC vector is known to be relatively independently with each others
  - ➢ The dynamic range of MFCC is vastly different across dimensions
  - ➢ The criterion decouples different dimensions, making the algorithm easier to develop and to implement.

Problem Formulation

- The solution is the conditional expectation

$$\hat{c}_x(k) = E\{c_x(k) \mid m_y\} = E\left\{\sum_b a_{k,b} \log m_x(b) \mid m_y\right\}$$
$$= \sum_b a_{k,b}(f) E\{\log m_x(b) \mid m_y\}$$

- Can be further simplified to

$$\hat{c}_x(k) \cong \sum_b a_{k,b}(f) E\{\log m_x(b) \mid m_y(b)\}$$

- The problem is reduce to finding the log-MMSE estimator of the Mel frequency filter bank's

$$\hat{m}_x(b) \cong \exp\left(E\{\log m_x(b) \mid m_y(b)\}\right)$$

Noise Suppressor for MFCC

- Set up a “straw man” by first rewriting

$$\hat{m}_x(b) = e\,\mathrm{xp}\left(E\left\{\log m_x(b) \mid m_y(b)\right\}\right) = e\,\mathrm{xp}\left(2E\left\{\log\sqrt{m_x(b)} \mid \sqrt{m_y(b)}\right\}\right)$$

- The same form in the objective function as the E&M log-MMSE amplitude spectral suppressor
- This naive approach has produced poor recognition results in our experiments.

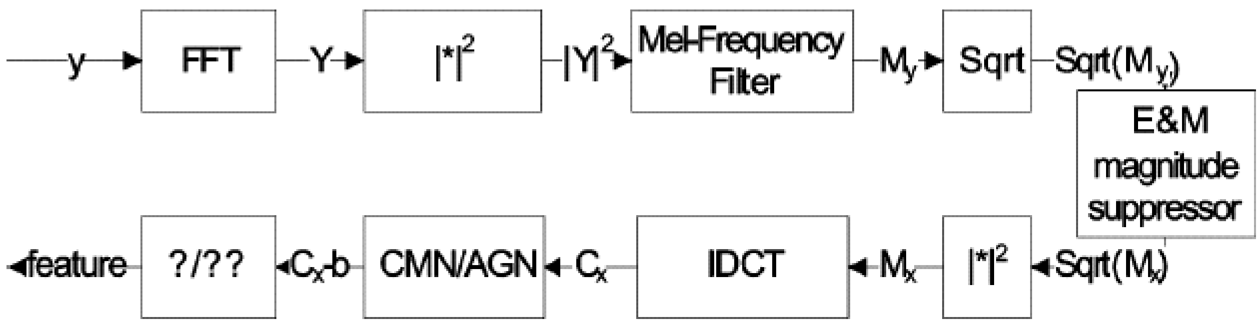

Fig. 1. Feature extraction pipeline where the E&M log-MMSE magnitude suppressor is directly applied to the magnitude spectrum of the filter bank output.

Noise Suppressor for MFCC

- Note that the filter bank' output $m_x(b)$, $m_n(b)$, and $m_y(b)$ take real value in the range of $(0,\infty]$, and thus it is inappropriate to model them with real-valued normal distributions.
- To develop appropriate models, we construct three artificial complex variable $M_x(b)$, $M_n(b)$, and $M_y(b)$ such that

$$|M_x(b)| = m_x(b) = \sum_f \omega_b(f)|X(f)|^2$$

$$|M_n(b)| = m_n(b) = \sum_f \omega_b(f)|N(f)|^2$$

$$|M_y(b)| = m_y(b) = \sum_f \omega_b(f)|Y(f)|^2$$

- We choose the ones with uniformly distributed random phases $\theta_x(b)$, $\theta_n(b)$, and $\theta_y(b)$.

Noise Suppressor for MFCC

- Since $M_y(b)$ contains all information there is in $m_y(b)$, can be rewritten as

$$\hat{m}_x(b) \cong \exp\left(E\left\{\log m_x(b) \mid M_y(b)\right\}\right)$$

- We follows the approach adopted in E&M by first evaluating the moment generating function

$$\Phi_b(\mu) = E\left\{\exp\left(\mu \log m_x(b) \mid M_y(b)\right)\right\}$$

$$= E\left\{m_x^u(b) \mid M_y(b)\right\}$$

$$\hat{m}_x(b) = \exp\left(\frac{d}{d\mu}\Phi_b(\mu)|_{\mu=0}\right) \qquad \frac{d}{d\mu} m_x^\mu = m_x^\mu \log m_x$$

Noise Suppressor for MFCC

- We assume that $\theta_x(b)$, $\theta_n(b)$, and $\theta_y(b)$ are independent and uniformly distributed random variables

$$\Phi_b(\mu) = E\left\{m_x^{\mu}(b) \mid M_y(b)\right\}$$
$$= \frac{\int_0^{\infty}\int_0^{2\pi} m_x^{\mu}(b)\, p\left(M_y(b), m_x(b), \theta_x(b)\right) dm_x(b)\, d\theta_x(b)}{p\left(M_y(b)\right)}$$
$$= \frac{\int_0^{\infty}\int_0^{2\pi} m_x^{\mu}(b)\, p\left(M_y(b) \mid m_x(b), \theta_x(b)\right) p\left(m_x(b), \theta_x(b)\right) dm_x(b)\, d\theta_x(b)}{\int_0^{\infty}\int_0^{2\pi} (b)\, p\left(M_y(b) \mid m_x(b), \theta_x(b)\right) p\left(m_x(b), \theta_x(b)\right) dm_x(b)\, d\theta_x(b)}$$

- $M_x(b)$ is assumed to follow the zero-mean complex normal distribution $p\left(m_x(b), \theta_x(b)\right) = \frac{m_x(b)}{\pi\sigma_x^2(b)} \exp\left\{-\frac{m_x^2(b)}{\sigma_x^2(b)}\right\}$
- Where

$$\sigma_x^2(b) \overset{def}{=} E\left\{\left|M_x(b)\right|^2\right\} = E\left\{m_x^2(b)\right\}$$

Noise Suppressor for MFCC

- Similarly, given that $M_y(b) - M_x(b)$

$$p\left(M_y(b) \mid m_x(b), \theta_x(b)\right) = \frac{1}{\pi\sigma_d^2(b)} \exp\left\{-\frac{\left|M_y(b) - m_x(b) e^{j\theta_x(b)}\right|^2}{\sigma_d^2(b)}\right\}$$

$$= \frac{1}{\pi\sigma_d^2(b)} \exp\left\{-\frac{\left|m_y(b) e^{j\theta_y(b)} - m_x(b) e^{j\theta_x(b)}\right|^2}{\sigma_d^2(b)}\right\}$$

$$= \frac{1}{\pi\sigma_d^2(b)} \exp\left\{-\frac{\left|m_y(b)\cos\left(\theta_y(b)\right) - m_x(b)\cos\left(\theta_x(b)\right) + j\left(m_y(b)\cos\left(\theta_y(b)\right) - m_x(b)\cos\left(\theta_x(b)\right)\right)\right|^2}{\sigma_d^2(b)}\right\}$$

$$= \frac{1}{\pi\sigma_d^2(b)} \exp\left\{-\frac{\left|m_y^2(b) + m_x^2(b) + 2m_y(b) m_x(b)\cos\left(\theta_y(b) - \theta_x(b)\right)\right|^2}{\sigma_d^2(b)}\right\}$$

$$where\ \sigma_d^2(b) \overset{def}{=} E\left\{\left|M_y(b) - M_x(b)\right|^2\right\} \geq E\left\{\left(m_y(b) - m_x(b)\right)^2\right\}$$

Noise Suppressor for MFCC

- Since $$m_y(b) = \sum_f \omega_b(f)\left|Y(f)\right|^2$$
$$= \sum_f \omega_b(f)\left(\left|X(f)\right|^2 + \left|N(f)\right|^2 + 2\left|X(f)\right|\left|N(f)\right|\cos\varphi(f)\right)$$
$$= m_x(b) + m_n(b) + \sum_f 2\omega_b(f)\left|X(f)\right|\left|N(f)\right|\cos\varphi(f)$$

  Where $\varphi(f)$ is the phase difference of X(f) and N(f)

$$\sigma_d^2(b) \geq E\left\{\left(m_n(b) + \sum_f 2\omega_b(f)\left|X(f)\right|\left|N(f)\right|\cos\varphi(f)\right)^2\right\}$$
$$= E\left\{m_n^2(b)\right\} + E\left\{\left(\sum_f 2\omega_b(f)\left|X(f)\right|\left|N(f)\right|\cos\varphi(f)\right)^2\right\}$$
$$where\ E\left\{2m_n(b)\left(\sum_f 2\omega_b(f)\left|X(f)\right|\left|N(f)\right|\cos\varphi(f)\right)\right\} \cong 0$$

$$\sigma_d^2(b) \cong \sigma_x^2(b) + \sigma_\varphi^2(b)$$

- One of major different from E&M. In E&M

$$\sigma_d^2(b) \overset{def}{=} E\left\{\left|Y(f) - X(f)\right|^2\right\} = E\left\{\left|N(f)\right|^2\right\} = \sigma_n^2(b)$$

Noise Suppressor for MFCC

- By substituting and replacing variable $\theta_y(b)-\theta_x(b)$ by $\beta(b)$

$$\Phi_b(\mu)=E\left\{m_x^{\mu}(b)\mid M_y(b)\right\}$$

$$=\frac{\int_0^\infty m_x^{\mu+1}(b)\exp\left\{-\frac{m_x^2(b)}{\sigma_x^2(b)}-\frac{m_x^2(b)}{\sigma_d^2(b)}\right\}g\left(m_x(b)\right)dm_x(b)}{\int_0^\infty m_x(b)\exp\left\{-\frac{m_x^2(b)}{\sigma_x^2(b)}-\frac{m_x^2(b)}{\sigma_d^2(b)}\right\}g\left(m_x(b)\right)dm_x(b)}$$

$$g\left(m_x(b)\right)=\int_0^{2\pi}\frac{1}{\pi\sigma_x^2(b)}\exp\left\{-\frac{2m_x(b)m_y(b)\cos\left(\beta(b)\right)}{\sigma_d^2(b)}\right\}d\beta(b)$$

- This can be show simplified

$$g\left(m_x(b)\right)=I_0\left(2m_x(b)\sqrt{\frac{v(b)}{\sigma^2(b)}}\right), where\ I_0(z)=\int_0^{2\pi}\exp\left(z\cos\beta\right)d\beta$$

$$\frac{1}{\sigma^2(b)}=\frac{1}{\sigma_d^2(b)}+\frac{1}{\sigma_x^2(b)}$$

$$v(b)=\frac{\xi(b)}{1+\xi(b)}\Upsilon(b)$$

Noise Suppressor for MFCC

- $v(b) = \frac{\xi(b)}{1+\xi(b)}\Upsilon(b)$ is defined from a priori signal-to-noise ratio

$$\xi(b) \overset{def}{=} \frac{\sigma_x^2(b)}{\sigma_d^2(b)} \cong \frac{\sigma_x^2(b)}{\sigma_n^2(b)+\sigma_\varphi^2(b)}$$

- And the adjusted a posteriori SNR

$$\Upsilon(b) \overset{def}{=} \frac{\sigma_y^2(b)}{\sigma_d^2(b)} \cong \frac{m_y^2(b)}{\sigma_n^2(b)+\sigma_\varphi^2(b)}$$

- Rewritten as

$$\begin{aligned}\Phi_b(\mu) &= E\left\{m_x^\mu(b) \mid M_y(b)\right\} \\ &= \frac{\int_0^\infty m_x^{\mu+1}(b)\exp\left\{-\frac{m_x^2(b)}{\sigma_x^2(b)}-\frac{m_x^2(b)}{\sigma_d^2(b)}\right\} I_0\left(2m_x(b)\right)\sqrt{\frac{v(b)}{\sigma^2(b)}}\,dm_x(b)}{\int_0^\infty m_x(b)\exp\left\{-\frac{m_x^2(b)}{\sigma_x^2(b)}-\frac{m_x^2(b)}{\sigma_d^2(b)}\right\} I_0\left(2m_x(b)\right)\sqrt{\frac{v(b)}{\sigma^2(b)}}\,dm_x(b)}\end{aligned}$$

Noise Suppressor for MFCC

$$\Phi_b(\mu) = \sigma^{\mu/2}\Gamma(\mu/2+1)M(\mu/2;1;-v(b))$$

$,$ where $\Gamma(\bullet)$ gamma function M(a;c;x) confluent hypergeometric function

$$\frac{\partial}{\partial\mu}M(\mu/2;1;-v(b))\bigg|_{\mu=0} = \frac{-1}{2}\sum_{r=1}^{\infty}\frac{(-v)^r}{r!}\frac{1}{r} \qquad \frac{\partial}{\partial\mu}\Gamma\left(\frac{\mu}{2}+1\right)\bigg|_{\mu=0} = \frac{-c}{2}$$

$$\frac{d}{d\mu}\Phi_b(\mu)\bigg|_{\mu=0} = \frac{1}{2}\ln\sigma + \frac{1}{2}\left(\ln v(b) + \int_{v(b)}^{\infty}\frac{(e)^{-t}}{t}dt\right)$$

$$\hat{m}_x(b) = \exp\left(E\{\log m_x(b) \mid m_y(b)\}\right) = G(\xi(b), v(b))m_y(b)$$

where

$$G(\xi(b), v(b)) = \frac{\xi(b)}{1+\xi(b)}\exp\left\{\frac{1}{2}\int_{v(b)}^{\infty}\frac{e^{-t}}{t}dt\right\}$$

- The MMSE estimate for MFCC is thus

$$\hat{c}(k) = \sum_b a_{k,b}E\{\log m_x(b) \mid m_y(b)\} = \sum_b a_{k,b}\log\left(G(\xi(b), v(b))m_y(b)\right)$$

Estimation of Parameters

- To apply the noise reduction algorithm, we need to estimate the noise variance $\sigma_n^2(b)$, the variance $\sigma_\varphi^2(b)$ and clean speech variance $\sigma_x^2(b)$
- Estimate of $\sigma_n^2(b)$
  - ➢ Using a minimum-controlled recursive movie-average noise tracker
  - ➢ A decision on whether a frame contains speech is made based on energy ratio test

$$\frac{\left|\ddot{m}_y(b)\right|_t^2}{\left|\ddot{m}_n(b)\right|_{\min}^2} > \vartheta$$

Where ϑ is threshold, $\left|\ddot{m}_n(b)\right|_{\min}^2$ is the smoothed minimum noise power, $\left|\ddot{m}_y(b)\right|_t^2$ is the smoothed power of the bth filter's output at the tth frame.

  - ➢ If the energy ratio is true the frame is assumed to contain speech the new noise estimate of the noise variance becomes

$$\sigma_n^2(b)_t = \sigma_n^2(b)_{t-1}, \text{ otherwise } \quad \sigma_n^2(b)_t = \alpha\sigma_n^2(b)_{t-1} + (1-\alpha)\left|m_y(b)\right|_t^2 \text{ using smoothing factor } \alpha$$

Estimation of Parameters

- Estimation of $\sigma_x^2(b)$
  - Using decision-directed approach.
  - $\sigma_x^2(b)$ for the current frame is estimate using the estimated clean speech from the previous frame and smoothed over the past frames.

Estimation of Parameters

- Estimation of $\sigma_\varphi^2(b)$

$$\sigma_\varphi^2(b) = E\left\{\left(\sum_f 2\omega_b(f)|X(f)||N(f)|\cos\varphi(f)\right)^2\right\}$$

$$= 4\sum_f E\left\{\left(\omega_b(f)|X(f)||N(f)|\cos\varphi(f)\right)^2\right\}$$

$$= 4\sum_f E\left\{\left(\omega_b(f)|X(f)||N(f)|\right)\right\}^2 \times \int_0^{2\pi} \frac{1}{2\pi}\cos\varphi(f)\,d\varphi(f)$$

$$= 2\sum_f E\left\{\left(\omega_b(f)|X(f)||N(f)|\right)^2\right\} = 2\sum_f \omega_b^2(f)\,E\left\{\left(|N(f)|\right)\right\}^2 E\left\{\left(|X(f)|\right)\right\}^2$$

  - Since we only estimate and keep track of statistics at the real-valued filter bank's output, we approximate $\sigma_\varphi^2(b)$ as

$$\sigma_\varphi^2(b) = 2\sum_f \omega_b^2(f)\,E\left\{\left(|N(f)|\right)\right\}^2 E\left\{\left(|X(f)|\right)\right\}^2$$

$$\cong 2E\left\{m_x(b)\right\}E\left\{m_n(b)\right\}\frac{\sum_f \omega_b^2(f)}{\sum_{f_1}\sum_{f_2}\omega_b(f_1)\omega_b(f_2)}$$

$$\cong 2\frac{E\left\{m_x(b)\right\}}{E\left\{m_n(b)\right\}}E\left\{m_n^2(b)\right\}\frac{\sum_f \omega_b^2(f)}{\left(\sum_f \omega_b(f)\right)^2}$$

$$\cong 2\frac{\sum_f \omega_b^2(f)}{\left(\sum_f \omega_b(f)\right)^2}\sqrt{\sigma_x^2(b)\,\sigma_n^2(b)}$$

Experiment setup

- Aurora3 corpus
- Close-talking or a hand-free microphone
- 39-dimension features used in our experiment
  - 13-dimension(with energy and without c0) static MFCC
  - Their delta and delta-delta feature
- The threshold ϑ was set to 0.9, and the parameter α set to 5.

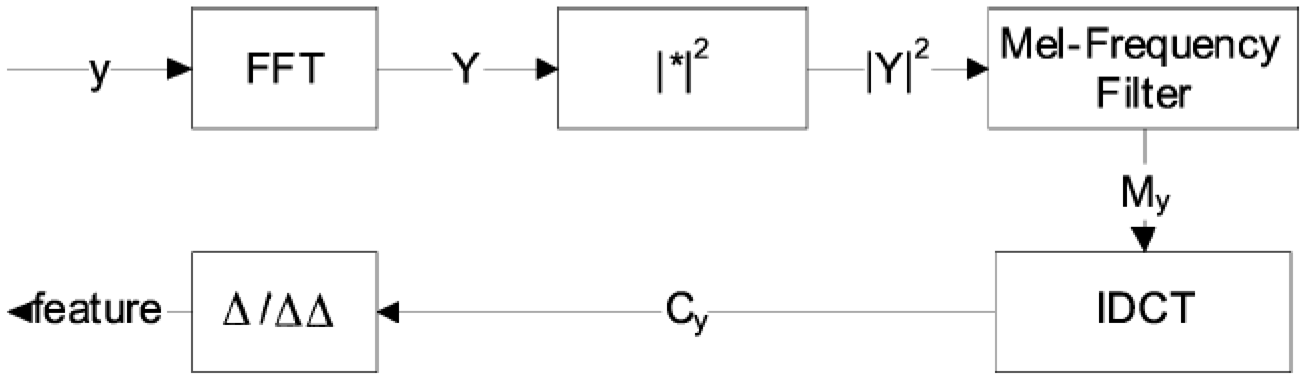

Fig. 5. Feature extraction pipeline for the ICSLP02 baseline system.

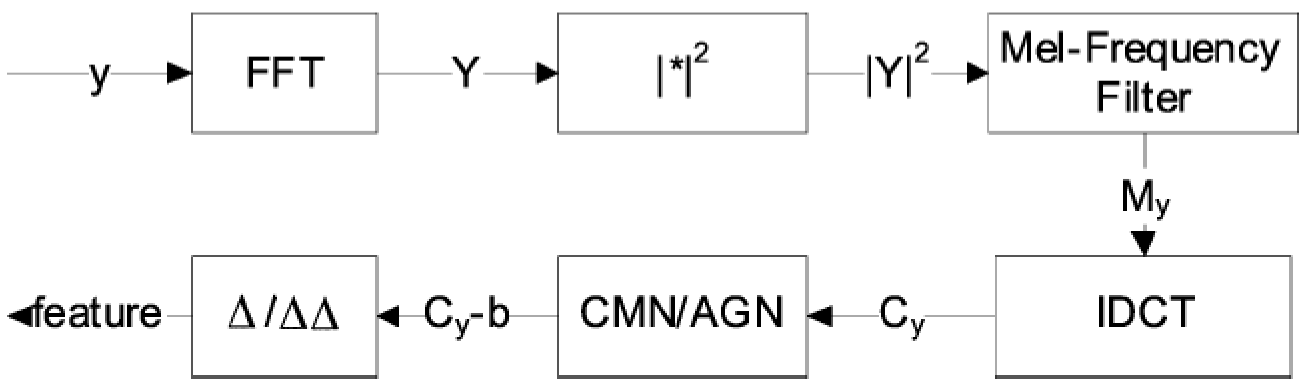

Fig. 6. Feature extraction pipeline for the CMN baseline system.

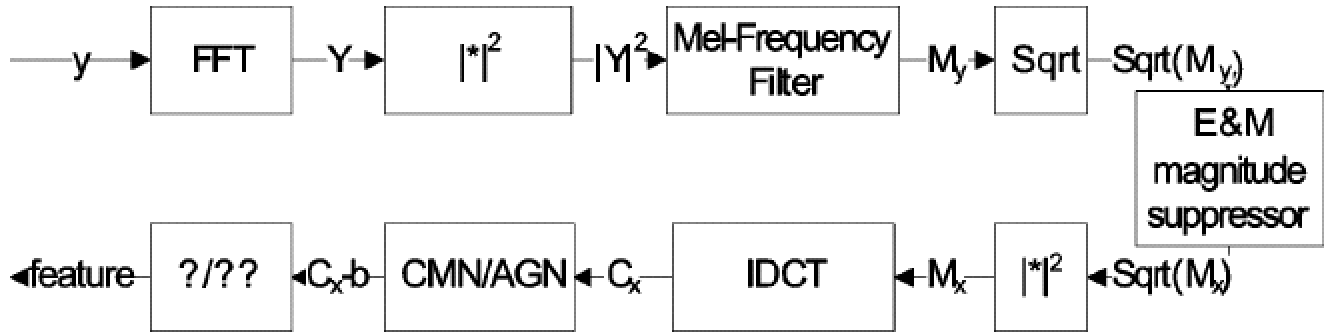

Fig. 1. Feature extraction pipeline where the E&M log-MMSE magnitude suppressor is directly applied to the magnitude spectrum of the filter bank output.

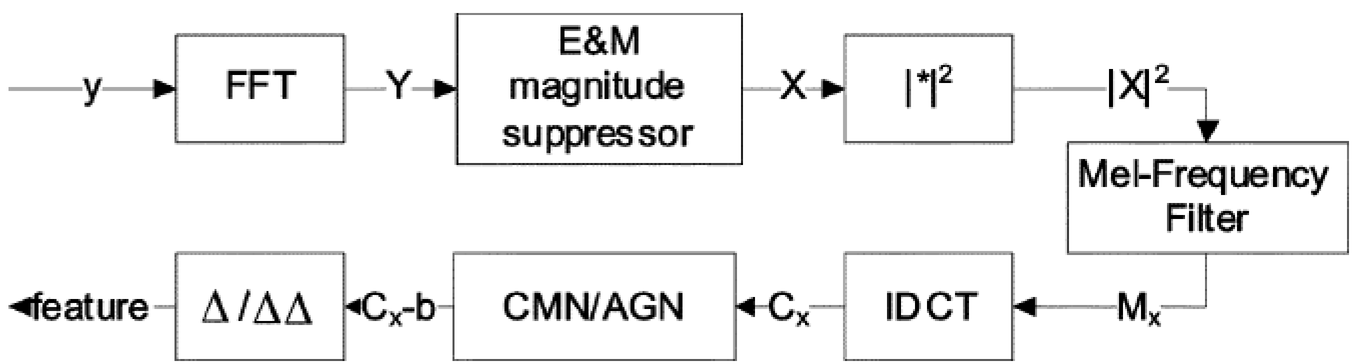

Fig. 7. Feature extraction pipeline for the E&M log-MMSE system [8], where the suppressor is applied to the DFT bins.

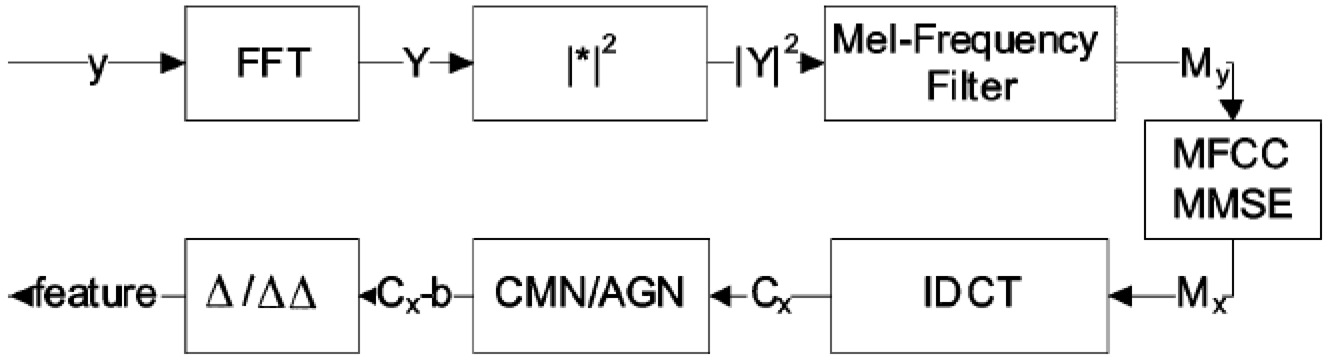

Fig. 8. Feature extraction pipeline for the MFCC-MMSE system.

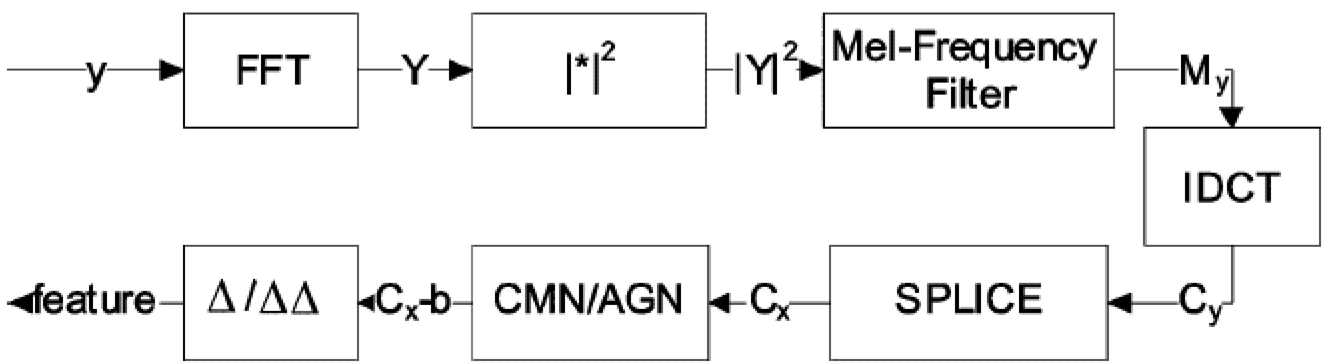

Fig. 9. Feature extraction pipeline for the SPLICE systems.

TABLE I

SUMMARY OF ABSOLUTE WER ON THE STANDARD TEST SETS IN THE AURORA-3 TASK UNDER DIFFERENT EXPERIMENTAL SETTINGS

Summary of Aurora 3 Absolute Word Error Rate (Standard Set)				
	Well	**Mid**	**High**	**Average**
ICSLP02 Baseline	8.96%	21.96%	48.85%	**23.48%**
CMN	6.87%	16.52%	31.11%	**16.31%**
FB Output Magnitude	6.87%	15.21%	31.29%	**15.89%**
E&M log-MMSE	5.57%	12.79%	29.23%	**14.01%**
MFCC-MMSE	5.08%	12.26%	23.26%	**12.13%**

TABLE II

SUMMARY OF RELATIVE WER REDUCTION ON THE STANDARD TEST SETS IN THE AURORA-3 TASK UNDER DIFFERENT EXPERIMENTAL SETTINGS

Summary of Aurora 3 Relative Improvement (Standard Set)			
Relative to →	**ICSLP02 Baseline**	**CMN**	**E&M log-MMSE**
CMN	30.55%	--	--
E&M log-MMSE	40.33%	14.08%	--
MFCC-MMSE	48.33%	25.59%	13.41%

TABLE III
DETAILED AURORA-3 ABSOLUTE WER RESULTS ON THE STANDARD TEST SETS UNDER THE MFCC-MMSE EXPERIMENTAL SETTING

Aurora-3 Word Error Rate with MFCC-MMSE (Standard Set)					
	Finnish	**Spanish**	**German**	**Danish**	**Average**
Well (x40%)	3.54%	5.90%	5.20%	5.66%	**5.08%**
Mid (x35%)	15.12%	5.39%	10.67%	17.84%	**12.26%**
High (x25%)	17.99%	34.77%	10.78%	29.49%	**23.26%**
Overall	**11.21%**	**12.94%**	**8.51%**	**15.88%**	**12.13%**

TABLE IV
DETAILED AURORA-3 WER REDUCTION RESULTS ON THE STANDARD TEST SETS AGAINST THE ICSLP02 BASELINE UNDER THE MFCC-MMSE

Aurora-3 Relative Improvement with MFCC-MMSE (Standard Set)					
	Finnish	**Spanish**	**German**	**Danish**	**Average**
Well (x40%)	51.24%	16.43%	40.91%	55.50%	**43.36%**
Mid (x35%)	22.42%	67.71%	43.72%	45.41%	**44.18%**
High (x25%)	69.75%	28.24%	59.82%	51.36%	**52.39%**
Overall	**54.44%**	**37.73%**	**49.54%**	**49.88%**	**48.32%**

TABLE VI

SUMMARY OF RELATIVE WER REDUCTION ON THE QUIET TEST SET IN THE AURORA-3 TASK UNDER DIFFERENT EXPERIMENTAL SETTINGS

Summary of Aurora 3 Relative Improvement (Quiet Set)		
Relative to ->	CMN	E&M log-MMSE
E&M log-MMSE	20.33%	--
MFCC-MMSE	21.72%	1.75%

TABLE VII

COMPARISON BETWEEN THE MFCC-MMSE SYSTEM AND THE ETSI'S AFE ON THE AURORA-3 TASK

Compare with AFE on Aurora 3 (Standard Set)			
	Well	Mid	High
ETSI AFE	4.70%	13.21%	12.75%
MFCC-MMSE	5.08%	12.26%	23.26%

TABLE VIII

COMPARISON BETWEEN THE MFCC-MMSE SYSTEM AND THE SPLICE ON AURORA-3 WHERE THE SPLICE CODE BOOK WAS TRAINED USING ADDITIONAL INFORMATION TO MAKE A MATCHING CONDITION

Comparisons with SPLICE on Aurora-3 (Standard Set)			
	Well	Mid	High
SPLICE	5.49%	13.55%	11.42%
MFCC-MMSE	5.08%	12.26%	23.26%