CROSS EVALUATION OF SPEECH ENHANCEMENT METHODS UNDER DIFFERENT NOISE CONDITIONS

Lara Nahma, Pei Chee Yong, Hai Huyen Dam, Sven Nordholm

Curtin University, Kent Street, Bentley, WA 6102, Australia
l.alibreesm@postgrad.curtin.edu.au
peichee.yong@nuheara.com { H.Dam, S.Nordholm }@curtin.edu.au

ABSTRACT

In this paper, we present a cross evaluation of different a priori SNR estimation methods as well as different time-frequency analysis processing using a subjective listening test. The noisy signal is corrupted by different types of background noise i.e. babble and pink, and varying levels of input SNR (0 dB and 10 dB) . The signals are processed using Short Time Fourier Transform (STFT) or Critical Band (CB) processing. After estimating the clean speech signal, it was presented to 10 participants for evaluation using a subjective listening test according to (ITU-TP.835) methodology. The results demonstrate that the participants preferred the speech signal processed using CB for low SNR levels and non-stationary background noise, which means that critical band based frequency scale is more useful in adverse noisy conditions.

***Index Terms*—speech enhancement, noise reduction, a priori SNR estimation, perceptually motivated speech enhancement, listening test.**

I. INTRODUCTION

The main challenge in single channel speech enhancement is to find the optimal denoising filter to reduce the background noise while preserving the speech components. In other words, the designed filter has to control the trade off between the noise reduction and speech distortion. Apart from that, it is also necessary to consider the quality of noise after suppression, since unnatural sounding background noise known as musical noise is bothersome for the listeners [1].

An important component in the speech enhancement system is the a priori SNR estimation [2], which involves an estimate of the clean speech and noise power spectral density (PSD). The state of the art decision directed (DD) based a priori SNR estimator is proposed by Ephraim and Malah [3]. This approach has the ability to reduce the annoying musical residual noise by lowering the variance of the a priori SNR estimate. However, the disadvantage of this approach is the slow adaptation towards speech onsets and offsets since its performance strongly depending on the a priori SNR estimate in the previous frame. This leads to a performance degradation in the speech enhancement system. To overcome this drawback, Yong et.al. [4] proposed a modified decision directed (MDD) a priori SNR estimator, which matches the current noisy speech spectrum with the a priori SNR estimate instead of the previous one.

One of the most important considerations that needs to be taken into account in the speech enhancement scheme is the speech characteristics. Since speech is highly non-stationary, dividing the degraded noisy signal into short frames is necessary in order to be able to treat the speech signal in each frame as approximately stationary. Spectral domain based speech enhancement utilizes the Short Time Fourier transform (STFT) for this purpose, [5], [6], [7].

The main limitation when using of STFT is that the analysis results in each frequency band has a uniform resolution which is not well adapted to the non-uniform resolution of the human auditory system [8]. Thus, using human auditory models as a pre-processor in speech enhancement system may improve the subjective quality and/or intelligibility for the enhanced speech [9].

Many proposals in the field of speech enhancement have represented the speech signals according to the human auditory system by applying auditory motivated filter bank like, Gammatone filter bank (GFB) [10], [11] bark scale based critical bands (CB) [12], [13]. Yet, it still unclear as to which time-frequency representation gives better performance in adverse environment, when the background noise level and characteristics are non-stationary. For that reason, a cross evaluation among different frequency resolution scales is important to investigate their impact on different speech estimators.

In this paper, we present a cross evaluation of a priori SNR estimators integrated with different time-frequency analysis techniques (STFT and CB) by a subjective listening test according to ITU-TP.835 methodology [7]. The subjective evaluation test demonstrates that the enhanced speech signals which were processed using CB achieved better results over those processed using STFT for low SNR levels and non-stationary background noise.

This paper is organized as follows. Section II describes single channel speech enhancement in STFT domain. Section III demonstrates the perceptually based speech enhancement system. Section IV presents the cross subjective quality evaluation. Section V, represents results and discussion, and Section VI concludes the paper.

II. SINGLE CHANNEL SPEECH ENHANCEMENT IN STFT DOMAIN

Let $s(t)$ and $v(t)$ denote speech and uncorrelated noise, respectively. The observed noisy speech signal $y(t)$ in the discrete time domain is given by

$$y(t) = s(t) + v(t) \tag{1}$$

Taking the STFT of the observed speech signal, we get

$$Y(k,m) = S(k,m) + V(k,m) \tag{2}$$

where $S(k,m)$ and $V(k,m)$ denote the complex spectral coefficients of speech signal and noise for a given frequency bin k and the time frame index m, respectively.

Figure 1 shows the framework of the single channel speech enhancement in STFT domain, where the clean speech estimate

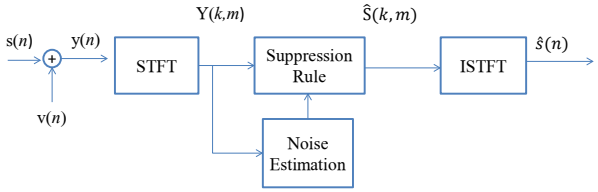

Fig. 1. Speech enhancement framework in STFT domain.

is obtained by applying a spectral magnitude weighting function $G_{\mathrm{STFT}}(k,m)$ to the noisy spectrum as given by

$$\hat{S}(k,m) = G_{\mathrm{STFT}}(k,m)Y(k,m) \tag{3}$$

The spectral magnitude weighting function often depends on the a posteriori SNR $\gamma(k,m)$

$$\gamma(k,m) = \frac{|Y(k,m)|^2}{\lambda_{\mathrm{v}}(k,m)} \tag{4}$$

and/or the a priori SNR $\xi(k,m)$

$$\xi(k,m) = \frac{\lambda_{\mathrm{s}}(k,m)}{\lambda_{\mathrm{v}}(k,m)} \tag{5}$$

where $\lambda_{\mathrm{s}}(k,m)$ and $\lambda_{\mathrm{v}}(k,m)$ represent the clean speech PSD and noise PSD, respectively.

In this work, we use the Wiener filter (WF) gain function [14], which is given by

$$G_{\mathrm{STFT}}(k,m) = \frac{\xi(k,m)}{1+\xi(k,m)}. \tag{6}$$

Finally, the speech estimate is obtained by taking the inverse STFT of the enhanced speech and using the overlap-add method

$$\hat{s}(n) = \mathrm{ISTFT}\left(\hat{S}(k,m)\right). \tag{7}$$

III. SINGLE CHANNEL SPEECH ENHANCEMENT IN CRITICAL BANDS

Since, the STFT has uniform resolution which is different from the natural filtering operation of the human auditory system, it is important to investigate if improvement in the speech perception can be obtained by using a time-frequency representation with non uniform resolution similar to the non linear frequency selectivity of the human ear. One simple way to achieve the human perceptual processing is by utilizing a critical band mapping from the STFT analysis of the noisy speech signal.

The block diagram for critical band speech processing is described in Figure 2. In the first step the noisy signal is transformed to the time-frequency domain by applying STFT with K frequency bins as in (2). In the second step, the output from the STFT $Y(k,m)$ is transformed into the critical band form by using an approximate analytical function to express the conversion from frequency f (in Hz) to critical band z (in bark scale), as given by [15]

$$f = 600\sinh\left(\frac{z}{6}\right). \tag{8}$$

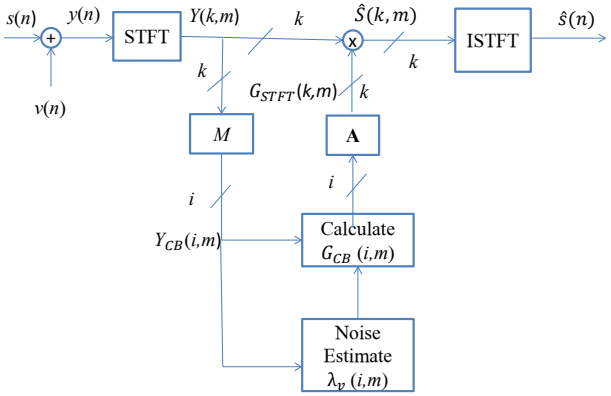

Fig. 2. Block diagram for the critical band processing.

By combining the FFT frequency bins into I critical bands, the noisy output in the critical band is expressed as follows

$$Y_{\mathrm{CB}}(i,m) = \sum_{k=1}^{K/2+1} M(i,k)\,|Y(k,m)| \tag{9}$$

where $i = [1,2,\cdots,I]$. The number of critical bands I is chosen according to the bark scale [15]. Here, $M(i,k)$ is the coefficient of critical bandpass filter which is defined by [16]

$$M(i,k) = \begin{cases} 10^{(z(k)-z_{\mathrm{c}}(i)+0.5)} & z(k) < z_{\mathrm{c}}(i) - 0.5 \\ 1 & z_{\mathrm{c}}(i)-0.5 < z(k) < z_{\mathrm{c}}(i)+0.5 \\ 10^{-2.5(z(k)-z_{\mathrm{c}}(i)-0.5)} & z(k) > z_{\mathrm{c}}(i)+0.5 \end{cases} \tag{10}$$

where $z_c(i)$ represents the center frequency of the i^{th} critical band. Then, the noisy spectrum $Y_{\mathrm{CB}}(i,m)$ is used to estimate the noise PSD $\lambda_{\mathrm{v}}(i,m)$, and a priori SNR estimation $\xi(i,m)$ for each band. Accordingly, the spectral magnitude weighting function $\mathbf{G}_{\mathrm{CB}}(m)$ in critical band for the m^{th} frame can be calculated as given by

$$\mathbf{G}_{\mathrm{CB}}(m) = [G_{\mathrm{CB}}(1,m), G_{\mathrm{CB}}(2,m), ..., G_{\mathrm{CB}}(I,m)]^T$$

$$G_{\mathrm{CB}}(i,m) = \frac{\xi(i,m)}{1+\xi(i,m)} \tag{11}$$

Once the weighting vector $\mathbf{G}_{\mathrm{CB}}(m)$ in critical band is calculated, it is interpolated back to the STFT resolution $\mathbf{G}(m)$ through an interpolation matrix $\mathbf{A}$,

$$\mathbf{G}_{\mathrm{STFT}}(m) = \mathbf{A}\mathbf{G}_{\mathrm{CB}}(m) \tag{12}$$

where the $\mathbf{A}$ matrix can be defined by least square approximation as $\mathbf{A} = (\mathbf{M}^T\mathbf{M})^{-1}\mathbf{M}^T$ and $\mathbf{M}$ represents the matrix with elements $M(i,k)$. From empirical findings, better results are obtained by simplifying the reconstruction matrix as

$$\mathbf{A} = \mathrm{diag}\left(\frac{1}{\mathbf{1M}}\right)\mathbf{M}^T \tag{13}$$

where $\mathbf{1}$ is $1 \times I$ row vector. The estimated speech in the STFT domain is then reconstructed by applying the interpolated weighting function $G_{\mathrm{STFT}}(k,m)$ on the noisy signal in (3).

IV. CROSS SUBJECTIVE QUALITY EVALUATION

In order to test the efficiency of the critical band based frequency scale compared to the conventional uniform scale (STFT), we present a cross evaluation of a priori SNR estimators integrated with different time-frequency analysis techniques (STFT and CB) by a subjective listening test according to ITU-T recommendation P.835 and conducted by ten participants. Three speech sentences consisting of 1 female speaker and 2 male speakers from the NOISEUS database [7] have been concatenated and corrupted with two different background noise (babble and pink) from NOISEX-92 database [17] for two levels of input SNR (0 dB and 10 dB). The noisy speech signals have been processed using three decision directed based a priori SNR estimation methods. The evaluated estimation methods include DD method [3], MDD method [4], critical band based a priori SNR estimation method (Prop-CB) [18] and STFT based a priori SNR estimation method (Prop-STFT) [19].

All the evaluated methods have been combined with Wiener filter (WF) gain function to estimate the clean speech signal. Minimum mean square error (MMSE) noise power estimator based on the speech presence probability [20] was employed to estimate the noise PSD for all the a priori SNR estimators. Once the clean speech signal estimate was obtained, it is presented to the participants of the subjective listening test.

The listening test was performed in a tranquil office room utilizing DT- 880 Beyerdynamic open air headphone. The test lasted around 25 minutes for each participant. Prior to giving their scores on the processed speech signals, the listeners were presented with the clean speech signal and the unprocessed speech signal as a kind of perspective for the best case and the worst case, individually. After that the participants were asked to listen and rate the enhanced signals according to ITU-T recommendation P.835. This methodology guides the participants to form the basis of their ratings regarding speech signal alone, background noise, musical noise and overall quality as shown in Table I.

Furthermore, to assess the difference between the listening test ratings, a statistical analysis of variance (ANOVA) is conducted to present a comparative analysis in reference to the unprocessed speech signal. A significant difference between scores was recognized depending on the obtained significance level (p-value). $\mathcal{H}$ represents the equality hypothesis and is defined as follows

$$\mathcal{H} = \begin{cases} \text{No significant difference is recognized,} & p > 0.05 \\ \text{Significant difference is recognized,} & p < 0.05 \end{cases} \quad (14)$$

which means if p > 0.05, equality hypothesis is accepted. Otherwise, the equality hypothesis is rejected.

V. RESULTS AND DISCUSSION

Figure 3 shows the mean results of the subjective listening test when the clean speech signal is corrupted by babble background noise for input SNR levels 0 dB and 10 dB. From the speech signals scores, it can be clearly observed that speech enhancement methods with CB have recorded higher scores than estimation methods with STFT at low SNR, while they achieve approximately same scores at high SNR. In terms of background noise, speech signals using CB have recorded higher scores than speech signals with STFT at low SNR, while in high SNR the background noise scores between CB and STFT are almost the same which means that all methods achieve the same amount of noise suppression.

Rating	Description
Speech	
5	very natural, no degradation
4	fairly natural, little degradation
3	somewhat natural, somewhat degraded
2	fairly unnatural, fairly degraded
1	very unnatural, very degraded
Background Noise	
5	not noticeable
4	somewhat noticeable
3	noticeable but not intrusive
2	fairly conspicuous, somewhat intrusive
1	very conspicuous, very intrusive
Musical noise	
5	not noticeable
4	somewhat noticeable
3	noticeable but not intrusive
2	fairly conspicuous, somewhat intrusive
1	very conspicuous, very intrusive

Table I: Scale description of the listening test criteria [4].

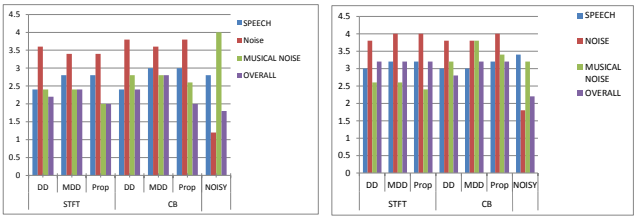

Fig. 3. Mean subjective listening test scores for speech processed by different speech enhancement methods and evaluated in babble background noise for two SNR levels (left side) 0 dB and (right side) 10 dB.

Musical noise results show that speech enhancement methods with CB achieved better results (higher musical noise scores) than those with STFT for different input SNR levels.

From the overall scores, it can be seen that the participants preferred speech signals with CB for low SNR levels. While at high SNR all methods achieves almost the same results.

Figure 4 shows the mean results of the subjective listening test when the clean speech signal is corrupted by pink background noise for input SNR levels 0 dB and 10 dB. In terms of musical noise results show that CB based methods have recorded the least amount of musical noise under different levels of input SNR values. Although listening test results for speech signal (speech scale) show that the participants preferred STFT enhanced speech signals over CB in pink noise condition at low SNR values, in high SNR (10 dB) the participants could not recognize any difference between STFT and CB based speech enhancement methods. Regarding background noise results (noise scale), it shows that better noise reduction is obtained in speech enhancement with STFT (higher noise scores) than CB methods under different SNR levels. From the overall scores, it can be observed that the participants preferred the signals estimated with STFT methods for low SNR values. In contrast, they preferred signals estimated with CB based methods at high SNR.

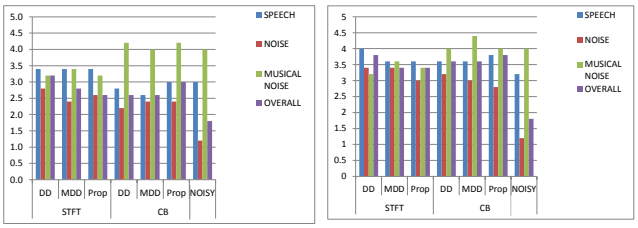

Fig. 4. Mean subjective listening test scores for speech processed by different speech enhancement methods and evaluated in pink background noise for two SNR levels (left side) 0 dB and (right side) 10 dB.

V-A. STATISTICAL ANALYSIS

Table II reports the obtained p-values of ANOVA test under different noise conditions. In terms of speech quality, The test shows that all obtained p-values are higher than 0.05, i.e., there was no statistically significant difference in speech quality between the obtained scores of the examined algorithms. This means that the enhanced speech signals did not contain a detectable speech distortion compared to the unprocessed speech signals. From background noise results, it can be clearly observed that all the estimation methods provided statistically significant differences when compared to the noisy speech signals in pink noise. However, in babble noise case there was no significant difference deemed for low SNR, whereas a significant difference achieved for high SNRs. In terms of musical noise, there was no significant difference between the enhanced speech signals that were estimated by the evaluated methods and the unprocessed speech signals detected in the different noise conditions and for varying levels of SNR. From the overall results it can be observed that only in the high SNR pink noise case, a significant difference was observed.

However, the above mentioned statistical analysis can not provide the answer as to which method performed better than the unprocessed speech signal. As such, along with ANOVA results, a post hoc comparison test according to Tukey's HSD test was also conducted to identify which method significantly improved the quality of the unprocessed speech signal. By comparing the scores obtained from the unprocessed speech signals and the scores obtained with speech signals enhanced by the various methods, the results of Tukey's HSD test are tabulated in Tables III. In this table asterisk indicates significant differences between enhanced speech signals and noisy signal. It can be observed that some methods only provided significant differences when compared to the unprocessed speech signal in terms of background noise and overall quality. In babble noise case, most of the SNR estimators for STFT and CB except (CB-Prop) achieved significant noise suppression over unprocessed speech signal at high SNR level. In pink noise case, speech signals estimated using (STFT-DD) and (STFT-Prop) methods achieved significant noise suppression compared to noisy signal for different levels of SNR. In contrast, the rest of the methods achieved better noise suppression performance than unprocessed speech signals in higher SNR level only. In terms of overall scale, the methods (DD-STFT and Prop-CB) significantly improved the overall quality when compared to the unprocessed speech signal for high SNR level.

Gain function	Noise	Input SNR	p-value			
			Speech	Background noise	Musical noise	Overall
WF	Babble	0dB	0.938	0.119	0.215	0.794
		10dB	0.984	0.014	0.460	0.416
	Pink	0dB	0.806	0.013	0.723	0.312
		10dB	0.235	0.0001	0.540	0.002

Table II: One way ANOVA test results to verify the statistically significant difference between different frequency warping scales used in the listening test under different noise conditions.

Noise	Input SNR	Rate	STFT			CB		
			DD	MDD	Prop	DD	MDD	Prop
Babble	0 dB	Speech						
		Noise						
		Musical noise						
		Overall						
	10 dB	Speech						
		Noise	*	*	*	*	*	
		Musical noise						
		Overall						
Pink	0 dB	Speech						
		Noise	*		*			
		Musical noise						
		Overall						
	10 dB	Speech						
		Noise	*	*	*	*	*	*
		Musical noise						
		Overall	*					*

Table III: Tukey's HSD Comparison between the enhanced speech signal using WF gain function and the unprocessed speech signal under different conditions.

VI. CONCLUSIONS

In this paper, a cross evaluation for STFT and CB processing was conducted by using subjective listening test. From the test results, it can be clearly noted that although the STFT method achieves better results in stationary background noise in terms of better noise suppression and speech quality, its performance degraded in non-stationary background noise and is generating more musical noise. On the other hand, CB processing provides significant benefit in terms of less musical noise under different noise conditions and different levels of input SNR. In addition, it achieves better performance compared to STFT in non-stationary noise (babble noise) especially for low SNR levels. This means that the proposed critical band based frequency scale is more useful in low SNR and non- stationary background noise. Further investigations are needed for a more complete set of noise conditions.

VII. REFERENCES

[1] S. Suhadi, C. Last, and T. Fingscheidt, "A data-driven approach to a priori snr estimation," *IEEE transactions on audio, speech, and language processing*, vol. 19, no. 1, pp. 186–195, 2011.

[2] C. Breithaupt and R. Martin, "Analysis of the decision-directed SNR estimator for speech enhancement with respect to low-SNR and transient conditions," *IEEE transactions on audio, speech, and language processing*, vol. 19, no. 2, pp. 277–289, 2011.

[3] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[4] P. C. Yong, S. Nordholm, and H. H. Dam, "Optimization and evaluation of sigmoid function with a priori snr estimate for real-time speech enhancement," *Speech Communication*, vol. 55, no. 2, pp. 358–376, 2013.

[5] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal processing*, vol. 81, no. 11, pp. 2403–2418, 2001.

[6] H. Gustafsson, S. E. Nordholm, and I. Claesson, "Spectral subtraction using reduced delay convolution and adaptive averaging," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 8, pp. 799–807, 2001.

[7] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2013.

[8] H. W. Löllmann and P. Vary, "Uniform and warped low delay filter-banks for speech enhancement," *Speech Communication*, vol. 49, no. 7-8, pp. 574–587, 2007.

[9] D. E. Tsoukalas, J. N. Mourjopoulos, and G. Kokkinakis, "Speech enhancement based on audible noise suppression," *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 6, pp. 497–514, 1997.

[10] S. Kortlang, S. D. Ewert, and T. Gerkmann, "Single channel noise reduction based on an auditory filterbank," in *proc. 14th International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2014, pp. 283–287.

[11] R. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, "An efficient auditory filterbank based on the gammatone function," in *a meeting of the IOC Speech Group on Auditory Modelling at RSRE*, vol. 2, no. 7, 1987.

[12] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," *IEEE Transactions on speech and audio processing*, vol. 7, no. 2, pp. 126–137, 1999.

[13] L. Singh and S. Sridharan, "Speech enhancement using critical band spectral subtraction." in *ICSLP*, 1998.

[14] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586–1604, 1979.

[15] A. Sekey and B. A. Hanson, "Improved 1-bark bandwidth auditory filter," *The Journal of the Acoustical Society of America*, vol. 75, no. 6, pp. 1902–1904, 1984.

[16] H. Hermansky, "Perceptual linear predictive (plp) analysis of speech," *the Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[17] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech communication*, vol. 12, no. 3, pp. 247–251, 1993.

[18] L. Nahma, P. C. Yong, H. H. Dam, and S. Nordholm, "Convex combination framework for a priori snr estimation in speech enhancement," in *2017 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, March 2017, pp. 4975–4979.

[19] ——, "Improved a priori snr estimation in speech enhancement," in *2017 23rd Asia-Pacific Conference on Communications (APCC)*, 2017, pp. 1–5.

[20] T. Gerkmann and R. C. Hendriks, "Noise power estimation based on the probability of speech presence," in *proc. IEEE Workshop on Applications of Signal Processing, Audio and Acoustics (WASPAA), New Paltz, NY*, 2011, pp. 145–148.