Binary Time-Frequency Mask for Improved Malay Speech Intelligibility at Low SNR Condition

Norezmi Jamal¹, N.Fuad¹,³, MNAH Sha’bani ¹,², Mohd Helmy Abd Wahab¹,³, Syed Zulkarnain Syed Idrus⁴

¹ Faculty of Electrical and Electronic, Universiti Tun Hussein Onn Malaysia 86400 Batu Pahat, Johor, Malaysia
² Centre for Diploma Studies, Universiti Tun Hussein Onn Malaysia 86400 Batu Pahat, Johor, Malaysia
³ Computational Signal, Imaging and Intelligence (CSII) Focus Group, Universiti Tun Hussein Onn Malaysia 86400 Batu Pahat, Johor, Malaysia
⁴ Faculty of Applied and Human Sciences, Universiti Malaysia Perlis, 01000 Kangar, Perlis, Malaysia

*Corresponding author’s e-mail: norfaiza@uthm.edu.my

Abstract. A binary masking is often seen as a medium to enhance speech signal especially in hearing aid and speech recognition applications due to its simplicity and efficient approach for supervised source separation. High intelligibility could be obtained by applying the binary time-frequency masking to noisy speech signal. Since the issue of linear filtering algorithms might affect the intelligibility of noisy speech signal, so, this paper presents the binary time-frequency mask for improved Malay speech intelligibility at low SNR condition. The clean Malay speech signals that were contaminated by car and train noise with different signal to noise ratio (SNR) and local criterion (LC) level in forming the binary mask are analysed. The performance of intelligibility improvement was evaluated using a normalized sub band enveloped correlation (nSec). Overall results showed that the proposed approach produces slightly improved speech intelligibility at low SNR value.

1. Introduction

Automatic speech recognition (ASR) is a great and powerful system that has been used in navigation and mobile phone application. However, during distant speech recognition, it is suffered by background noise that may lead to misrecognition of spoken words or sentences. As a result, quality and intelligibility of recognized speech may be affected. Speech quality is defined as how much intensity of speech signal could be perceived which is highly subjective while speech intelligibility is a measure of how recognizable and meaningful speech is in any given distortions [1]. To overcome this issue, speech enhancement algorithm is required at front end of ASR system [2]. Basically, the purpose of speech enhancement algorithms may include improving the quality of the speech signal and increasing the intelligibility as well [3]. Unfortunately, improving speech intelligibility is more challenging compared to speech quality [3, 4]. This is because of speech intelligibility is mostly depend on the content of the speech material. The requirement of intelligibility enhancement by
suppressing the background noise without distorting the underlying target speech signal is important [5].

Several researchers carried out research on speech enhancement algorithms for monaural or single microphone to enhance the speech signal [6]. These algorithms have shown to improve speech quality but their capability of increasing the intelligibility of speech in noise has remained somewhat elusive [7] especially during low signal to noise ratio (SNR) [8]. There are several factors that contribute to the lack of intelligibility improvement in conventional noise reduction algorithms [7]. For example, it is often not possible to estimate accurately enough the background noise spectrum especially in non-stationary types of noise (e.g. multi-talker babble) [9]. Next, the distortions introduced by the frequency specific gain functions can at times be more damaging than the background noise itself [9] and lastly, removal of low intensity speech sounds such as unvoiced consonants which is important for intelligibility [9]. Based on aforementioned limitations, binary masking algorithm approach such as Ideal Binary Mask (IBM) has been recently applied in speech recognition applications [9, 10] rather than noise reduction algorithm [3, 11] in order to achieve higher intelligibility purpose.

Theoretically, target speech signal is separated by applying IBM to noisy speech signal. The idea of IBM algorithm is to improve the overall Signal to Noise Ratio (SNR), which the weighted sum of SNRs across all bands of frequency by getting rid of frequency bands with low SNRs while remaining the ones with high SNRs [12]. The advantage of this approach was discussed in the intelligibility enhancement evaluation studies, which showed good performance on synthesized speech using IBM based algorithm [9, 13-15]. Most of researchers used the open source available database for analysis of enhanced speech especially with English spoken language [14, 16, 17]. But, for Malay database, not much works has been explored yet. For example, authors in [18] conducted research on Malay speech enhancement with high SNR rather than low SNR noisy speech.

Thus, the aims of this paper are to suppress noise without introducing any perceptible distortion and to improve speech intelligibility on different types of background noisy conditions for Malay spoken language using binary time-frequency approach. The following paper is organised as follows: (a) Section 2 describes the theoretical study of binary time-frequency mask algorithm; Section 3 presents the research method; Section 4 shows the experimental results for binary time-frequency mask with different local criterion (LC) and SNR, followed by Section 5 which summarises the findings.

2. Binary Time-frequency Binary Mask

In this paper, binary time-frequency mask is known as ideal binary masking (IBM). It is a concept that was formulated based on the perceptual auditory masking phenomenon by reducing the audibility of a weaker sound in the presence of a higher intensity sound signal within the same critical band. The IBM is a technique that decomposes the input signal into time-frequency (T-F) units and classifies them into target-dominated or masker-dominated regions by a predefined threshold called local signal to noise ratio criterion (LC). The IBM algorithm, \( B(k, t) \) is constructed as [3, 15]:

\[
B(k, t) = \begin{cases} 
1 & \text{if } SNR(k, t) > LC \\
0 & \text{otherwise}, 
\end{cases}
\]  

(1)

where \( SNR(k, t) \) denotes the target speech energy at time-frame, \( t \) and channel \( k \) in the time-frequency (T-F) unit, which is measured in decibel. The choice of local criterion is not unique, and when applied to corrupted speech spectra, can produce large gains in speech intelligibility [3]. So, \( B(k, t) \) is assigned as value of 0 to a T-F unit if local SNR is not greater than LC, and 1 otherwise.

3. Methodology

Figure 1 shows a binary time-frequency (T-F) mask workflow in this paper. The target speech signals are Malay spoken language by a female, including two Malay phrases such as “Selamat
“Petang” and “Selamat Pagi”. Noted that the recording session were done in clean environment room with size of 6.5 x 7.5 feet using SL300 condenser microphone at sampling rate is 16 kHz. While car and train noise signal are acquired from NOIZEUS database with sampling rate of 8 kHz. All speech signals were resampled from 16 kHz to 8 kHz sampling rate in order to combine with noise signal as to ensure similar sampling rate for generation of noisy speech signal is achieved. For synthesize noisy speech signal, IBM algorithm is applied as shown in Equation 1.

A binary mask was formed by comparing the local Signal to Noise Ratio (SNR) within a T-F unit against local criterion (LC), by assigning 1 if the local SNR was greater than LC and 0 otherwise. The synthesis of noisy speech was done after converting raw noisy speech signal into T-F domain with low-pass FIR filtering using 20 ms of Hanning window and binary mask is applied. Lastly, the performance of IBM algorithm is measured using objective evaluation such as Normalized Subband Envelope Correlation (nSec) to measure gain in intelligibility enhancement. The equation is [19]

\[
\text{nSec} = \sum \sum \frac{T(\tau, k)Y(\tau, k)}{\|T(\tau, k)\|\|Y(\tau, k)\|}
\]

\[\text{(2)}\]

![Figure 1: Binary time-frequency mask workflow](image)

4. Results and discussion

Figure 2 (i) and Figure 2 (ii) illustrates the spectrograms of Malay phrase for “Selamat Petang” and “Selamat Pagi”, respectively during clean, noisy and synthesized speech condition. For clean speech spectrogram, it shows very low intensity for sound of /s/, which is known as fricatives sound [3] that extremely difficult to differentiate from background noise compared to other sound classes. While Malay vowels like /e/ and /a/ are relatively high in intensity due to darker color of stripe in spectrogram, which easier to be determined as well as more intelligible. In short, most of consonants such as /p/ and /t/ sounds are also overlapping with the background environment. On the other hand, both Malay phrases are difficult to be determined during noisy environment at low SNR, especially at -10 dB.
Figure 2: Spectrograms of synthesized Malay speech via the IBM Mask technique at SNR of -10dB

For ideal binary mask (IBM Mask) spectrogram, black regions indicate time-frequency elements that are dominated by the speech. Next, both synthesized speech signals are observed approximately similar to target clean speech. However, the performance of binary masking is difficult to evaluate based on observation of spectrogram. Due to that, normalized subband envelope correlation (nSec) value was calculated. Intelligibility enhancement performance of both Malay phrases with different types of noise signals for different LC parameters, by varying the SNR from -25dB to 5 dB were obtained in Figure 3 and Figure 4, respectively. The curves plotted show that the measured normalized subband envelope correlation (nSec) value is lower at low SNR value compared to higher SNR value. This is because of binary masking approach is difficult to be applied at very low SNR value and much depend on threshold of LC value. nSec is higher with low LC value during car noise type, which is up to -25 dB. While for train noise type, LC value is much higher at low SNR value.

As seen in Figure 3(i) and Figure 3(ii), the local criterion (LC) that makes intelligibility performance at high normalized subband envelope correlation (nSec) is between -25 dB SNR and 0 dB SNR for inside car noisy condition, which produced nSec is greater than 70% at optimal LC value for both Malay phrases. While Figure 4(i) and Figure 4(ii) depicts that nSec for both Malay phrases is greater than 80% at optimal LC value for different SNR value during inside train noisy condition. The choice of LC value is essential to produce higher nSec as well as to preserve enough speech information when LC value is smaller than SNR value of the mixture [8]. Overall, it shows that for masker signal with different LC value, the IBM produces slightly improved average performance at low SNR value. The cross-correlation coefficient is used to measure the similarity between internal representations of the target and enhanced speech signal. Unfortunately, this approach is very challenging to realize as an automatic system in human auditory system when keep varying the value of LC parameters manually [20].
5. Conclusion
The intelligibility of noisy Malay speech signals has been successfully improved by applying a binary time-frequency mask for different types of background noise. This study of binary mask by classifying each frame of time-frequency as "speech-dominated" or "noise-dominated" based on local
criterion (LC) parameter condition have been determined. Even though normalized subband envelope correlation (nSec) are able to improve speech intelligibility at low Signal to Noise Ratio (SNR) level, but this approach is suffering from keep varying the value of LC parameter at optimal value and not suitable for automatic system during distant speech recognition. In order to improve this approach, it is recommended to propose Artificial Neural Network based mask estimation for supervised Malay speech separation from mixture signal.

6. References

[1] J. Boldt, "Binary masking & speech intelligibility," Ph. D. thesis, Aalborg Universitet, 2011.
[2] N. Jamal, S. Shanta, F. Mahmud, and M. Sha’abani, "Automatic speech recognition (ASR) based approach for speech therapy of aphasics: A review," in *AIP Conference Proceedings*, 2017, vol. 1883, no. 1, p. 020028: AIP Publishing LLC.
[3] P. C. Loizou, *Speech enhancement: theory and practice.* CRC press, 2013.
[4] H. Kaur and R. Talwar, "Analysis of the SNR Estimator for Speech Enhancement Using a Cascaded Linear Model," *International Journal of Advanced Computer Science and Applications*, vol. 7, 2016.
[5] A. Sulong, T. S. Gunawan, O. O. Khalifa, M. Kartiwi, H. J. I. J. o. E. Dao, and C. Engineering, "Single Channel Speech Enhancement using Wiener Filter and Compressive Sensing," vol. 7, no. 4, p. 1941, 2017.
[6] N. Saleem and M. I. Khatkhat, "A review of supervised learning algorithms for single channel speech enhancement," *International Journal of Speech Technology*, vol. 22, no. 4, pp. 1051-1075, 2019.
[7] G. Kim and P. C. Loizou, "Gain-induced speech distortions and the absence of intelligibility benefit with existing noise-reduction algorithms," *The Journal of the Acoustical Society of America*, vol. 130, no. 3, pp. 1581-1596, 2011.
[8] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 22, no. 12, pp. 1849-1858, 2014.
[9] L. Lightburn, E. De Sena, A. Moore, P. A. Naylor, and M. Brookes, "Improving the perceptual quality of ideal binary masked speech," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 661-665: IEEE.
[10] J. B. Boldt, U. Kjems, M. S. Pedersen, T. Lunner, and D. Wang, "Estimation of the ideal binary mask using directional systems," in *Proceedings of the 11th International Workshop on Acoustic Echo and Noise Control*, 2008.
[11] E. Verteletskaya and B. Simak, "Enhanced spectral subtraction method for noise reduction with minimal speech distortion," in *Proceedings of the 17th International Conference on Systems, Signals and Image Processing*, 2010, pp. 61-64.
[12] D. Wang, U. Kjems, M. S. Pedersen, J. B. Boldt, and T. Lunner, "Speech intelligibility in background noise with ideal binary time-frequency masking," *The Journal of the Acoustical Society of America*, vol. 125, no. 4, pp. 2336-2347, 2009.
[13] N. Li and P. C. Loizou, "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction," *The Journal of the Acoustical Society of America*, vol. 123, no. 3, pp. 1673-1682, 2008.
[14] U. Kjems, J. B. Boldt, M. S. Pedersen, T. Lunner, and D. Wang, "Role of mask pattern in intelligibility of ideal binary-masked noisy speech," *The Journal of the Acoustical Society of America*, vol. 126, no. 3, pp. 1415-1426, 2009.
[15] J. Chen and D. Wang, "DNN based mask estimation for supervised speech separation," in *Audio source separation*: Springer, 2018, pp. 207-235.
[16] Y. Li and D. Wang, "On the optimality of ideal binary time–frequency masks," *Speech Communication*, vol. 51, no. 3, pp. 230-239, 2009.
[17] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125-2136, 2011.

[18] A. Hussain and S. Z. M. S. Mukari, "Single channel speech enhancement using ideal binary mask technique based on computational auditory scene analysis," *Journal of Theoretical and Applied Information Technology*, vol. 91, no. 1, pp. 12-22, 2016.

[19] J. B. Boldt and D. P. Ellis, "A simple correlation-based model of intelligibility for nonlinear speech enhancement and separation," in *2009 17th European Signal Processing Conference*, 2009, pp. 1849-1853: IEEE.

[20] T. Bentsen, T. May, A. A. Kressner, and T. Dau, "The benefit of combining a deep neural network architecture with ideal ratio mask estimation in computational speech segregation to improve speech intelligibility," *PloS one*, vol. 13, no. 5, 2018.