Improving the Performance of Speech Perception in Noisy Environment Based on a FAME Strategy

Abstract

Background noise is a significant factor influencing the performance of speech perception. Previous studies showed that the temporal fine structure (TFS) plays an important role in speech perception, and that frequency amplitude modulation encoding (FAME) is a successful approach to enhance the TFS information for cochlear implant (CI) recipients. Following the success of FAME for CI recipients, this study aims to evaluate the speech perception performance of FAME for normal hearing (NH) listeners in noisy conditions. Experimental results from the present study confirm that FAME provides better speech perception performance and lower listening effort for NH listeners. In particular, FAME improved Mandarin disyllabic words recognition by as much as 16.7 percentage points and the ease of listening by 1.6 (MOS scale). This demonstrates that the FAME strategy is promising for improving speech recognition performance for NH listeners in noisy environment.

Index Terms: speech perception, frequency amplitude modulation encoding (FAME), noise reduction.

1. Introduction

Speech perception is typically good under quiet listening conditions for normal hearing (NH) listeners. However, many common environmental factors such as noise negatively affect speech understanding [2], especially for older adults [3]. Noise is present in an acoustic environment and it masks the speech signal by obscuring the less intense portions of the signal (e.g., consonant parts) [4]. The result is a reduction in the redundancy of acoustic and linguistic cues in speech, and this effect increases as the signal-to-noise ratio (SNR) decreases.

Unsupervised noise reduction (NR) is a common approach to assist human hearing under noisy conditions. Successful techniques include the log minimum mean squared error (logMMSE) [5], Karhunen-Loéve transform (KLT) [6, 7], and generalized maximum a posteriori spectral amplitude (GMAPA) [8] methods. Advances have also been made to develop NR algorithms that suppress noise without introducing much distortion to the speech signal [9]. The evaluation criterion of most of the NR approaches is SNR improvement of speech distortion; however, these have been shown to improve primarily the subjective quality of speech rather than speech intelligibility [10, 11]. Speech quality is highly subjective in nature and can be easily improved by suppressing the background noise. On the other hand, intelligibility is related to the content of the spoken words and can be improved only by suppressing the background noise without damaging the speech signal. Designing such an unsupervised NR approach to improve intelligibility in noisy conditions has been extremely challenging, because of unreliable estimates of background noise signal from the corrupted noisy speech. Therefore, existing unsupervised NR approaches can improve speech quality but not speech intelligibility [12].

It has been found that the temporal fine structure (TFS) [13, 14] information of speech plays an important role for speech perception under noisy conditions [13, 15], especially for hearing impaired peoples. Nie et al. [1] proposed a frequency amplitude modulation encoding (FAME) strategy to evaluate the potential contribution of TFS information to speech recognition in noisy conditions via acoustic simulations of cochlear implants (CI) recipients (detail of the FAME technology can be found in section 2). The results showed that, for sentence recognition in the presence of a competing voice, noisy speech processed by FAME can improve performance as much as 71 percentage points for CI recipients. This implies a potentially large benefit of providing (or emphasizing) TFS information for the performance of speech perception for NH listeners. Following the study of [1] assessing the contribution of FAME for speech understanding in noise for CI recipients, the present work further examined the effect of the FAME strategy on speech perception for NH people in noisy conditions.

The remainder of the paper is organized into five sections. The following section reviews the FAME strategy. The remaining sections will focus on the experiment design, and results and discussion. Finally, our findings are summarized in Section 5.

2. Frequency amplitude modulation encoding (FAME)

The FAME strategy [1] was proposed in 2005 to encode both amplitude and frequency modulations in order to improve CI performance in noisy conditions. It transforms the fast-varying TFS into slowly varying frequency modulation (FM) signals that are applied to the carrier in each band. Fig. 1 shows the structure of the FAME strategy, which contains an analysis part and a synthesis part. First, the speech signal is divided into K bands by a bank of bandpass filters (bands 1 to K). The amplitude modulation (AM) and frequency modulation (FM) signals are extracted in separate pathways in each band.

Fig. 2(a) shows extraction of AM in the K-th subband. The AM is extracted by full-wave rectification of the output of the bandpass filter, followed by a low-pass filter (LPF 1) to obtain a slowly varying amplitude modulation signal AM_K AM. The purpose of LPF 1 is to control the maximal AM rate preserved in the AM signal. Finally, delay compensation is introduced to synchronize the amplitude and frequency modulation pathways.

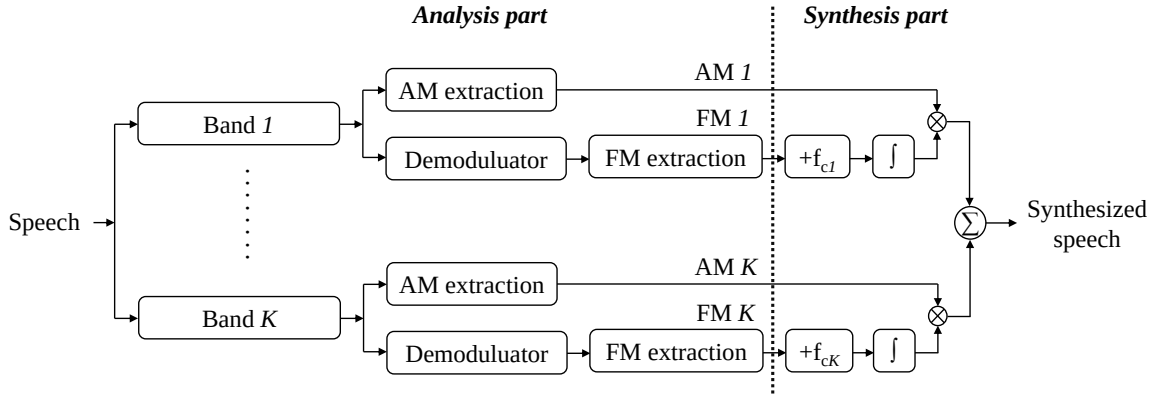

Fig. 1. Structure of the FAME strategy [1].

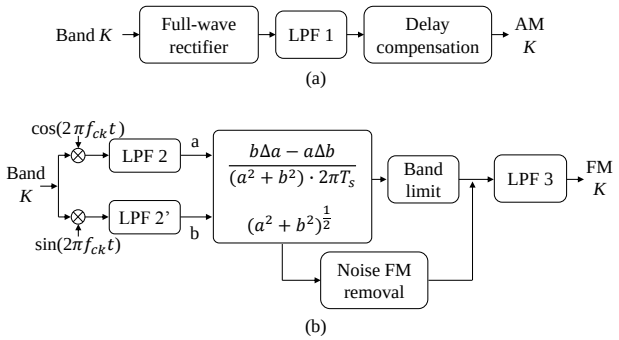

Fig. 2. Block diagram for extracting (a) AM and (b) FM in the FAME strategy. [1]

Fig. 2(b) shows the FM extraction pathway. First, the output *K*-th band, $x_K(t)$, is subjected to a quadrature oscillator with the center frequency (f_{cK}). This operation is equivalent to shifting the spectrum of $x(K)$ from f_{cK} to zero and $2\,f_{cK}$ in the frequency domain. Next, the slowly varying frequency components (i.e., a and b) are extracted by low-pass filters (LPF 2 and LPF 2'). Note that the cutoff frequencies of LPF 2 and LPF 2' determine FM depth and bandwidth. These slowly varying frequency components can be divided into two parts: in-phase and out-of-phase signals. The in-phase signal can be filtered out by the low-pass filter LPF2 to produce:

$$\mathrm{a} = \tfrac{1}{2} m(t) cos\varphi(t). \tag{1}$$

Similarly, the out-of-phase signal can be filtered out by the low-pass filter LPF2' to produce:

$$\mathrm{b} = -\tfrac{1}{2} m(t) sin\varphi(t) = \tfrac{1}{2} m(t) cos[\varphi(t) + \pi/2]. \tag{2}$$

Dividing the out-of-phase signal (2) by the in-phase signal (1) will produce:

$$\begin{aligned} \frac{b}{a} &= -\tan\varphi(t) \\ \varphi(t) &= \tan^{-1}\left(-\frac{b}{a}\right). \end{aligned} \tag{3}$$

Finally, the instantaneous frequency can be obtained:

$$\mathrm{FM} = \tfrac{1}{2\pi} \cdot \frac{d\varphi(t)}{dt} = \frac{d\tan^{-1}\left(-\frac{b}{a}\right)}{2\pi dt} = \frac{b(da/dt) - a(db/dt)}{2\pi(a^2+b^2)}. \tag{4}$$

In discrete implementation, differentiation in (4) can be substituted by calculating the difference in time (Δ) to obtain the slowly varying frequency modulation:

$$\mathrm{FM} = \frac{b\Delta a - a\Delta b}{2\pi(a^2+b^2)\times T_s}. \tag{5}$$

Where T_s represents the sampling period. In addition, an additional amplitude calculation, Eq. (6), is used as a threshold device to remove erroneous frequency modulation produced by low-level noise owing to the differential process in FM extraction. Finally, a low-pass filter (LPF 3) is used to produce the desired slowly varying FM signals. Note that the cutoff frequency of LPF 3 determines the FM rate. More detailed information on FAME can found in [1].

$$\text{Noise FM removal} = \left[(a^2 + b^2)^{1/2}\right]. \tag{6}$$

3. Experiments

3.1. Subjects and materials

Twelve (18–22 yrs., 6 female) NH native-Mandarin speakers participated in the listening experiment. Mandarin disyllabic words lists from [16] and the Taiwan Mandarin version of Hearing in Noise Test (TMHINT) lists from [17] were used as the testing materials to test the speech intelligibility and listening effort [18]. All sentences were pronounced by the Hidden Markov Models (HMM)-based Speech Synthesis System [19], and recorded at a sampling rate of 16 kHz. Train noise was used as the masker signal to corrupt test sentences at three signal-to-noise ratio (SNR) levels of -5, 0, and 5 dB, which were chosen to avoid the ceiling/floor effects.

3.2. Procedure

The parameters of *K*, LPF1, LPF2, LPF2', and LPF3 in FAME were set to 16, 500 Hz, 200 Hz, 200 Hz, and 400 Hz, respectively. We conducted two sets of experiments, namely

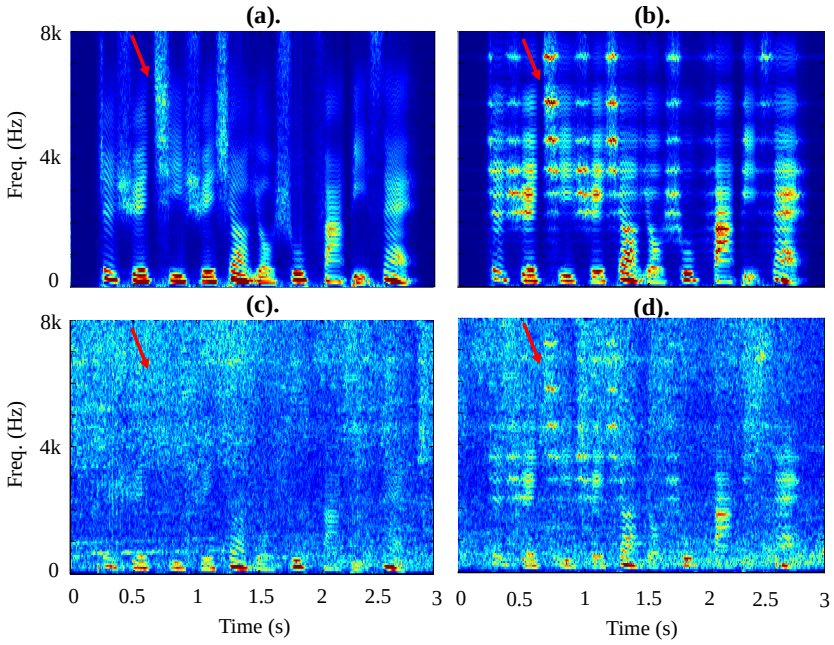

Fig. 3. Spectrograms of (a) clean speech, (b) clean_FAME speech, (c) clean speech corrupted by train noise (SNR level= 0 dB), and (d) clean_FAME speech corrupted by train noise (SNR level= 0 dB). The sentence was a synthesis by the Hidden Markov Models (HMM)-based speech synthesis system in Mandarin, saying "There is a calligraphy competition in this semester".

objective and subject listening evaluations. For the objective listening test, we adopted the confusion disyllabic word lists [20], and automatically selected a set based on a genetic algorithm [16]. It is a closed set (i.e., multiple choice question) speech intelligibility testing approach. Each subject participated in a total of six [= 2 signal processed approaches × 3 SNR levels × 1 masker] testing conditions. Each condition contained 15 disyllabic words. The order of the six conditions was randomized across subjects, and none of the 15 disyllabic words were repeated across testing conditions. Subjects were requested to choose the answer based on what they heard, and allowed to repeat the stimuli twice.

In the subjective listening tests, we adopted a questionnaire [21] to investigate the listening effort between original and FAME approaches. The questions for the effort ratings were "How much effort was required for you to identify the components of the sentence?" There were three SNR levels (-5, 0, and 5 dB) were used. Each subject participated in a total of six [= 2 signal processed approaches × 3 SNR levels × 1 masker] testing conditions. Each condition contained 10 sentences (10 words). Each subject was asked to score these statements for each of the noisy signals using a mean opinion score (MOS) comparison. The MOS rating is the most widely used measure for subjective quality tests in which the subjects rate the test speech scale from 5 (no effort) to 1 (extremely high effort).

4. Results

4.1. Spectrogram analyses

A spectrogram shows the spectral representations of a time-varying signal and is often used to analyze frequency and level properties of speech signals [22]. Figure 3 illustrates four sub-figures, showing the spectrograms of: (a) clean speech; (b) clean speech processed by FAME, denoted as clean_FAME; (c) clean speech corrupted by train noise (SNR level = 0 dB); and (d) clean_FAME speech corrupted by train noise (SNR level = 0 dB). The sentence used in Fig. 3 was synthesized by a HMM-based speech synthesis system [19] in Mandarin, saying "*There is a calligraphy competition in this semester*". From the example of Fig. 3 (c), we can note that the speech information of high frequency parts (i.e., 3 kHz to 8 kHz) will be affected by the background noise. In other words, the high frequency information of speech will not be easy to hear for the subjects. Therefore, a lot of consonant information (high frequency parts) will be lost, thus decreasing the performance of speech perception, especially for Mandarin listening [23]. Compared with the clean speech in Fig. 3, the clean_FAME speech provided a sturdier high frequency speech signal. From Fig. 3 (d), it can be noted that the FAME strategy not only maintains the speech information of the low frequency parts, but also further emphasizes high frequency speech signals. Therefore, it can help subjects to hear consonant information of Mandarin well in noisy conditions to improve the performance of speech intelligibility.

4.2. Listening test

4.2.1. Speech intelligibility

The average scores of speech intelligibility between original and FAME-processed speech at three different SNR levels are shown in Fig. 4. The average speech perception scores for {original and FAME} are {65.0% and 81.7%} at -5 dB, {73.3% and 84.5%} at 0 dB, and {84.4% and 88.3%} at 5 dB, respectively. To verify the improvement, we conducted a paired sample *t*-test. Original vs. FAME resulted in a p-value < 0.001, denoting significant improvement. The results confirm the

effectiveness of FAME in improving the intelligibility of enhanced speech.

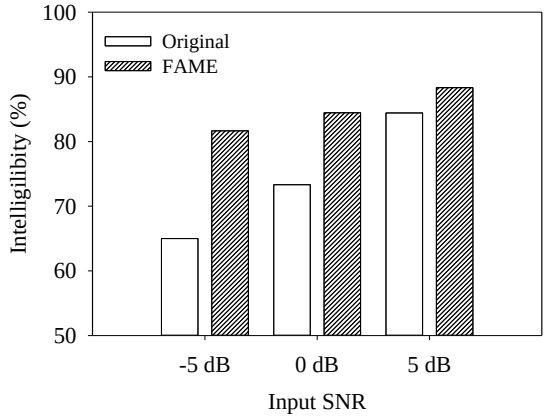

Fig. 4 Average scores of speech perception scores at different SNR levels between original and FAME processed

4.2.2. Listening effort

Listening effort [18] is used to quantify the cognitive resources necessary for speech understanding in noisy conditions [24]. That is, increased listening effort means more brainpower is needed to recognize and understand speech. In this study, the subjects rated listening effort on a five-point rating scale that ranged from no effort to extremely high effort. A higher score indicates that the processed approach was preferred. The subjects gave ratings for the two processing conditions (i.e., original and FAME signal processed approach) at the -5, 0, and 5 dB SNR levels, thus six different conditions. The order of the six test conditions was randomized across subjects. Each test condition contained 10 sentences.

Fig. 5 shows the average listening effort (MOS rating) of 12 subjects for all test conditions. The average listening effort scores for {original and FAME} were {1.41 and 3.01} at -5 dB, {2.28 and 3.73} at 0 dB, and {3.14 and 4.08} at 5 dB, respectively. The paired sample *t* tests confirm that listening effort scores are significantly different ($p < 0.001$) between original and FAME processing.

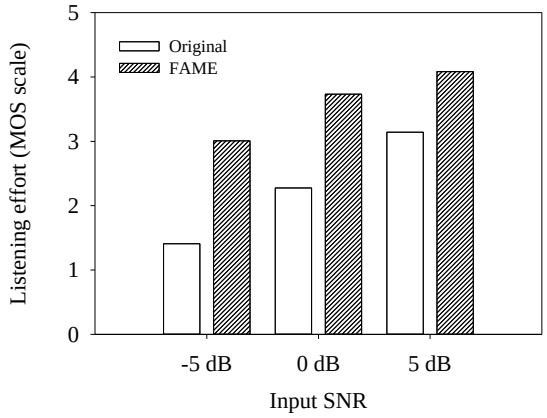

Fig. 5 Average results of listening effort over 12 subjects for different input SNRs between original and FAME processed.

5. Conclusions

Following the study of [1] assessing the contribution of the FAME strategy for speech understanding of CI recipients in noisy conditions, the present work further examined its effect for NH listeners. The FAME strategy can enhance and centralize the TFS information, especially for consonants (i.e., high frequency sounds), to improve the performance of speech perception. Fig. 3 showed an example where the FAME strategy could provide stronger consonant energy than original speech for NH listeners. In other words, the FAME technique can strengthen the high frequency part of speech, such that it is not easily affected by noise. The consonant information is one of the key factors for speech intelligibility in Mandarin. The results of listening tests showed that FAME provided higher intelligibility and lower listening effort for NH listeners. These results demonstrate that the FAME strategy is a highly promising technique to enhance speech understanding in noisy conditions for NH listeners.

6. Acknowledgements

7. References

[1] K. Nie, G. Stickney, and F. G. Zeng, "Encoding frequency modulation to improve cochlear implant performance in noise," *Biomedical Engineering, IEEE Transactions on,* vol. 52, no. 1, pp. 64-73, 2005.

[2] C. C. Crandell, and J. J. Smaldino, "Classroom acoustics for children with normal hearing and with hearing impairment," *Language, speech, and hearing services in schools,* vol. 31, no. 4, pp. 362-370, 2000.

[3] S. Anderson, A. Parbery-Clark, T. White-Schwoch, and N. Kraus, "Auditory brainstem response to complex sounds predicts self-reported speech-in-noise performance," *Journal of Speech, Language, and Hearing Research,* vol. 56, no. 1, pp. 31-43, 2013.

[4] K. S. Helfer, and L. A. Wilber, "Hearing loss, aging, and speech perception in reverberation and noise," *Journal of Speech, Language, and Hearing Research,* vol. 33, no. 1, pp. 149-155, 1990.

[5] Y. Ephraim, and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *Acoustics, Speech and Signal Processing, IEEE Transactions on,* vol. 33, no. 2, pp. 443-445, 1985.

[6] A. Rezayee, and S. Gazor, "An adaptive KLT approach for speech enhancement," *Speech and Audio Processing, IEEE Transactions on,* vol. 9, no. 2, pp. 87-95, 2001.

[7] Y. Hu, and P. C. Loizou, "A generalized subspace approach for enhancing speech corrupted by colored noise," *Speech and Audio Processing, IEEE Transactions on,* vol. 11, no. 4, pp. 334-341, 2003.

[8] Y. Tsao, and Y. H. Lai, "Generalized maximum a posteriori spectral amplitude estimation for speech enhancement," *Speech Communication,* 2015.

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

[10] Y. Hu, and P. C. Loizou, "A comparative intelligibility study of single-microphone noise reduction algorithms,"

The Journal of the Acoustical Society of America, vol. 122, no. 3, pp. 1777-1786, 2007.
[11] P. Loizou, "Subjective evaluation and comparison of speech enhancement algorithms," *Speech Commun*, vol. 49, pp. 588-601, 2007.
[12] P. C. Loizou, and G. Kim, "Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 1, pp. 47-56, 2011.
[13] B. C. Moore, "The role of temporal fine structure processing in pitch perception, masking, and speech perception for normal-hearing and hearing-impaired people," *Journal of the Association for Research in Otolaryngology*, vol. 9, no. 4, pp. 399-406, 2008.
[14] F. Chen, Y. Tsao, and Y.-H. Lai, "Modeling speech intelligibility with recovered envelope from temporal fine structure stimulus," *Speech Communication*, vol. 81, pp. 120-128, 2016.
[15] F. G. Zeng, K. Nie, G. S. Stickney, Y.-Y. Kong, M. Vongphoe, A. Bhargave, C. Wei, and K. Cao, "Speech recognition with amplitude and frequency modulations," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 7, pp. 2293-2298, 2005.
[16] P. C. Li, Y. Y. Chiang, K. S. Tsai, and S. T. Young, "Genetic algorithm for the efficient selection of disyllabic word lists used in Mandarin speech discrimination tests," *Medical and Biological Engineering and Computing*, vol. 43, no. 5, pp. 648-657, 2005.
[17] M. W. Huang, "Development of Taiwan Manadarin hearing in noise test," Department of speech language pathology and audiology, National Taipei University of Nursing and Health science, 2005.
[18] I. Brons, R. Houben, and W. A. Dreschler, "Effects of Noise Reduction on Speech Intelligibility, Perceived Listening Effort, and Personal Preference in Hearing-Impaired Listeners," *Trends in hearing*, vol. 18, pp. 2331216514553924, 2014.
[19] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, "Speech synthesis based on hidden Markov models," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1234-1252, 2013.
[20] S. S. Young, *All-in-One solution for hearing-impaired rehibitition*, NSC 102-2218-E-715 -001, 2013.
[21] P. A. Gosselin, and J.-P. Gagne, "Older adults expend more listening effort than young adults recognizing speech in noise," *Journal of Speech, Language, and Hearing Research*, vol. 54, no. 3, pp. 944-958, 2011.
[22] S. Haykin, *Advances in spectrum analysis and array processing (vol. III)*: Prentice-Hall, Inc., 1995.
[23] Y. H. Lai, T. C. Liu, P. C. Li, W. T. Shih, and S. T. Young, "Development and Preliminary Verification of a Mandarin-Based Hearing-Aid Fitting Strategy," *PLOS ONE*, vol. 8, no. 11, pp. e80831, 2013.
[24] C. B. Hicks, and A. M. Tharpe, "Listening effort and fatigue in school-age children with and without hearing loss," *Journal of Speech, Language, and Hearing Research*, vol. 45, no. 3, pp. 573-584, 2002.