JASA ARTICLE

Auditory filter-bank compression improves estimation of signal-to-noise ratio for speech in noise

Fangqi Liu,[1,a] Andreas Demosthenous,[1] and Ifat Yasin[2]
[1]Department of Electronic and Electrical Engineering, University College London, London WC1E 7JE, United Kingdom
[2]Department of Computer Science, University College London, London WC1E 6BT, United Kingdom

ABSTRACT:
Signal-to-noise ratio (SNR) estimation is necessary for many speech processing applications often challenged by nonstationary noise. The authors have previously demonstrated that the variance of spectral entropy (VSE) is a reliable estimate of SNR in nonstationary noise. Based on pre-estimated VSE-SNR relationship functions, the SNR of unseen acoustic environments can be estimated from the measured VSE. This study predicts that introducing a compressive function based on cochlear processing will increase the stability of the pre-estimated VSE-SNR relationship functions. This study demonstrates that calculating the VSE based on a nonlinear filter-bank, simulating cochlear compression, reduces the VSE-based SNR estimation errors. VSE-SNR relationship functions were estimated using speech tokens presented in babble noise comprised of different numbers of speakers. Results showed that the coefficient of determination (R^2) of the estimated VSE-SNR relationship functions have absolute percentage improvements of over 26% when using a filter-bank with a compressive function, compared to when using a linear filter-bank without compression. In 2-talker babble noise, the estimation accuracy is more than 3 dB better than other published methods. © *2020 Acoustical Society of America*. https://doi.org/10.1121/10.0001168

(Received 13 July 2019; revised 14 March 2020; accepted 9 April 2020; published online 5 May 2020)
[Editor: Michael I. Mandel] Pages: 3197–3208

I. INTRODUCTION

Signal-to-noise ratio (SNR) is a measure that compares the level of the desired signal (S) to the level of background noise (N), the ratio of signal power to noise power. Knowledge of the degree of background noise corruption is necessary for optimizing signal processing strategies in many acoustic applications, such as speech enhancement and automatic speech recognition. In the case of speech enhancement, the aim is to reduce the background additive noise without reducing speech intelligibility. Most speech enhancement algorithms accomplish this by applying a gain function (Gerkmann and Hendriks, 2012) based on certain error criteria (e.g., mean square error), to multiply the magnitude spectrum of speech which is corrupted by background noise (noisy-speech). Such an enhancement strategy relies particularly on accurate estimates of the SNR, as the gain function itself or the optimization of the gain function often depend on the estimated SNR (Plapous *et al.*, 2006).

Generally, the SNR can be estimated over either a long time-scale (often greater than 1000 ms, referred to here as global SNR), or a short time scale, with duration of less than 30 ms (referred to here as instantaneous SNR). Instantaneous SNR is often preferred in conventional speech enhancement algorithms (Ephraim and Malah, 1984; Plapous *et al.*, 2006). This is because the instantaneous SNR tracks the rapid fluctuations of the noise power more closely, and this benefits speech enhancement in nonstationary noise. However, in practice, the estimated instantaneous SNR in nonstationary noise can be erroneous due to the rapid fluctuations of noise power. These errors can cause speech distortion in speech enhancement (Loizou and Kim, 2011). In some cases, global SNR estimation can provide a more accurate estimation of SNR compared to an instantaneous SNR estimation, in both stationary and nonstationary noise (May *et al.*, 2017; Papadopoulos *et al.*, 2016). This is thought to be because the longer time-scale noise power is more stable than that using a shorter time scale. Moreover, regulating the underlying gain function based on a longer time scale can lead to a smoother noise reduction. Martin *et al.* (2004) reduced the speech distortion in a minimum-mean-square-error based enhancement algorithm by regulating the lower limit of the gain adjustment according to the global SNR. Currently, there is increased interest in using a global SNR estimate to optimize conventional speech enhancement algorithms (Martin *et al.*, 2004) or to develop supervised speech enhancement algorithms (Healy *et al.*, 2013) for improved speech intelligibility.

The accuracy of global SNR estimation is challenged by nonstationary noise, for instance, babble noise, which is one of the most common types of interfering background noise. Babble noise is composed of multiple talking speakers. The estimation of the SNR in babble noise is often challenging because the statistics of babble noise are similar to that of clean speech that varies considerably over time (Krishnamurthy and Hansen, 2009). Vondrášek and Pollák (2005) estimated the SNR of noisy-speech by estimating noise power using a "hard decision" based voice activity

[a]Electronic mail: fangqi.liu.14@ucl.ac.uk, ORCID: 0000-0002-0903-7462.

detection (VAD). The "hard decision" approach analyses the speech using time-windowing short frames, and then decides which frames contain or do not contain speech. The decision criteria are often based on the assumption that the power in a frame that contains speech (speech-present frame) will be higher than that of a frame that does not contain speech (speech-absent frame). This is because the speech-present frame contains both noise and clean-speech, while the speech-absent frame contains noise only. However, since the noise power is only updated when speech is absent, when speech presents itself, the estimation of noise power is delayed. The estimation accuracy of this method decreases when there is a sudden rise of noise power during speech-present frame evaluation. The noise power tracking delay can be reduced using a "soft decision" approach, which uses *a priori* SNR to decide the degree of speech absence, and updates the noise power even during the speech present evaluation phase (Gerkmann and Hendriks, 2012). However, in highly nonstationary noise (e.g., babble noise with fewer talkers), the noise power may still fluctuate within the reduced delay time of noise power estimation.

Kim and Stern (2008) took a different approach and compared the amplitude distributions of clean speech (speech uncorrupted by background noise) and noise. They assumed that clean-speech amplitude can be described by a Gamma distribution, while the noise amplitude can be described by a Gaussian distribution. They found that the parameter of noisy-speech amplitude distribution can be used to estimate SNR. The relationship function between the distribution parameter and the SNR is assessed to estimate the SNR according to the measured distribution parameter of the noisy-speech. Since the distribution parameter is measured without detecting speech presence, in comparison to the noise power estimation method (Gerkmann and Hendriks, 2012), the noise power tracking delay is avoided. However, the SNR estimation accuracy of noisy-speech could be severely degraded in babble noise (Narayanan and Wang, 2012), since the amplitude distribution of babble noise is similar to that of clean-speech.

To accurately estimate SNR in babble noise, we proposed using a measure called the variance of spectral entropy (VSE) (Liu *et al.*, 2017). VSE is defined as the variance (over time) of the spectral entropy of noisy-speech. Similar to the method used in Kim and Stern (2008), we estimate the relationship between VSE and SNR and save it in a lookup table. The SNRs of noisy-speech are estimated by measuring the VSE of noisy-speech. Spectral entropy was first used in VAD, and has shown to be more robust in nonstationary noise (Wu and Wang, 2005) because spectral entropy is independent of the amount of noise power; it is robust against the fluctuations of nonstationary noise power. Moreover, VSE characterizes the signal variability as it measures the spectral entropy variation over time. Measuring signal variability has been shown to be more robust in nonstationary noise (Ghosh *et al.*, 2011). In contrast to the long term signal variability (LTSV) (Ghosh *et al.*, 2011), which calculates the variance (over the spectrum) of time-domain entropy for hundreds of fast Fourier transform (FFT) frequency bins, VSE increases computational efficiency by calculating the variance (over time) of frequency-domain entropy from the output of an auditory filter-bank with a small number of frequency bands. In a previous approach (Liu *et al.*, 2017), the SNR estimation accuracy showed apparent degradation in highly nonstationary noise (e.g., 2-talker babble noise). The accuracy of the VSE based SNR estimation method relies on the stability of the VSE at a given SNR level. In highly nonstationary noise, the time and frequency characteristics of noisy speech vary rapidly. Consequently, the accuracy of the VSE estimation and the VSE-SNR relationship function are degraded. The reduced accuracy of SNR estimation in highly non-stationary noise may have occurred because the VSE was calculated using a linear auditory filter-bank (Liu *et al.*, 2017).

It is known that the response of the cochlea increases with increasing sound stimulus level, but the response growth is compressive (compression). At moderate and high levels, the input/output function of the cochlea has a compression exponent (rate of change, measured in dB/dB) of less than 1 (Ruggero *et al.*, 1997). Aspects of cochlear compression have been successfully applied in contemporary hearing assistive devices for restoring audibility and comfortable loudness growth. The compressed gain extends the hearing dynamic range by applying greater gain to low level signals and less gain to high level signals. As a result, the spectral contrasts of both clean-speech and noise are reduced (Moore *et al.*, 1998). In this paper, we implement a nonlinear filter-bank, which simulates the compressive response of cochlea. The compressive function may reduce the variation of the VSE, especially in highly non-stationary noise situations. In the case of VSE-based SNR estimation, the reduced spectral contrasts would reduce the spectral differences over noisy-speech samples, and reduce the variation of the VSE over different noisy-speech samples at the same SNR. The current study applied a nonlinear auditory filter-bank to calculate the VSE, which was in turn used to estimate the SNR. The proposed approach is evaluated in nonstationary noise conditions when the interfering background is babble noise containing different numbers of talkers. The coefficient of determination (R^2) of the VSE-SNR relationship function to random generated noisy-speech samples is evaluated. The SNR estimation accuracy of the compression-based approach is compared with that of the waveform amplitude distribution analysis (WADA) method (Kim and Stern, 2008), noise power estimated (NPE) based method (Gerkmann and Hendriks, 2012), minimum mean-square error (MMSE) based clean speech estimation method (Erkelens *et al.*, 2007), and our previous approach using VSE with a linear filterbank (Liu *et al.*, 2017).

II. THEORY OF VSE BASED SNR ESTIMATION

Information entropy, which is defined by the negative logarithm of the probability of each given data, was first

https://doi.org/10.1121/10.0001168

described by Shannon (1948) to characterize the amount of information produced by a stochastic source based data. In acoustic signals, the entropy is often calculated over the spectrum. By further calculating the VSE over time, the signal variability is characterized. The signal variability reflects the degree of nonstationarity of the signal that can be used to track the SNR, even in nonstationary noise. This is because clean-speech has information encoded in both frequency and time domains that has higher signal variability than most environmental (broadband) background noise (Ghosh *et al.*, 2011). For example, a person with an average speaking rate produces approximately 10–15 phonemes with different spectral characteristics per second (Liberman, 1996). Although interfering nonstationary noise could have higher variability, it often has a spectrum flatter than that of clean-speech. The resultant VSE estimated for nonstationary noise would be distinctive from the VSE estimated for clean-speech. The noise corruption degrades the signal variability of clean-speech, and the degree of noise corruption (SNR) can be expressed as a function of VSE, which has been demonstrated in previous work (Liu *et al.*, 2017).

A. VSE calculation

We model the noisy-speech signal $y(i)$ as the sum of a clean-speech signal $x(i)$ with the corrupting background noise $d(i)$,

$$y(i) = x(i) + d(i), \tag{1}$$

where i denotes the sampling time index. In the present study, we made the following assumptions: First, clean speech and noise are statistically independent across time and frequency. Second, the speech and noise amplitude across time have a mean value of zero. To calculate the VSE, the instantaneous spectral entropy needs to be first calculated. In contrast to conventional methods (Shen *et al.*, 1998; Wu and Wang, 2005) that use the derived probability associated with spectral energy of hundreds of FFT frequency bins, we use the probability associated with the instantaneous power of the signal in each filter-bank frequency band to reduce computational complexity. Although our calculated spectral entropy has a lower spectral resolution, we have previously (Liu *et al.*, 2017) demonstrated that an auditory filter bank comprised of ten frequency bands is sufficient to acquire an almost linear VSE-SNR relationship. In fact, a higher spectral resolution might degrade the stability of the VSE-SNR relationship function because a higher spectral resolution increases the degree of freedom of the spectral entropy. The probability $p(k,i)$ for frequency band k at the sampling time i is calculated by normalizing the instantaneous spectral power across all frequency bands (Wu and Wang, 2005),

$$p(k,i) = \frac{S(k,i)}{\sum_{l=1}^{K} S(l,i)} \quad k = 1 \ldots K. \tag{2}$$

$S(k,i)$ is the instantaneous power of signal in frequency band k,

$$S(k,i) = |g(k)(H(k,i) * y(i))|^2. \tag{3}$$

$H(\mathrm{k},\mathrm{i})$ is the transfer function of the band pass filter k. g is the gain applied after filtering (referring to the gain of the filter-bank; for a linear filter-bank $g = 1$ dB/dB, 1 dB/dB refers to the growth rate of the gain function, which is the gain increase by 1 dB per dB; see Ruggero *et al.*, 1997, for more details of representing gain in decibel scale). K is the total number of frequency bands. According to Eq. (6) used in Wu and Wang (2005), the spectral entropy $h(i)$ is calculated by

$$h(i) = -\sum_{k=1}^{K} [p(k,i)\log(p(k,i))]. \tag{4}$$

$$\text{The VSE is: } \sigma_H(j) = \frac{1}{M}\sum_{i=1}^{M} (h_j(i) - \bar{h}_j)^2, \tag{5}$$

where $\bar{h}_j$ is the mean value of spectral entropy (MSpE) in each SNR estimation interval j. M is the total number of the sampling points over the estimation time interval. In this study, the duration of SNR estimation interval was chosen to be 1000 ms.

B. Analysis of the VSE-SNR relationship

To analyse the factors that influence the VSE-based SNR estimate accuracy, we derive the relationship function between VSE and SNR. Let W denote the number of sample points containing speech with added noise (speech presence) over an interval j. The number of sample points only containing noise (when speech is absent) is $M - W$. Assuming that the spectral entropy of noise and clean-speech are independent, Eq. (5) can be rewritten as

$$\sigma_H(j) = \frac{1}{M}\left\{\sum_{i=1}^{M-W} (h_D(i) - \bar{h}_j)^2 + \sum_{i=1}^{W} (h_Y(i) - \bar{h}_j)^2\right\}, \tag{6}$$

where $h_D(i)$ is the spectral entropy of the sample points only containing noise and $h_Y(i)$ is the spectral entropy of points containing speech added with noise. For simplification, the estimation time interval index j will be omitted in the following equations. We assume estimations of spectral entropy at each time interval are independent. MSpE ($\bar{h}_j$) in Eq. (6) can be expressed as a function of the MSpE when speech is absent [$\bar{h}_D^{M-W}(j)$] and when speech is present [$\bar{h}_Y^{W}(j)$]. Thus,

$$\sigma_H = \frac{M-W}{M}\sigma_{hd}^{M-W} + \frac{W}{M}\sigma_{hy}^{W}\left(1 - \frac{W}{M}\right)\left(\bar{h}_D^{M-W} - \bar{h}_Y^{W}\right)^2, \tag{7}$$

where σ_{hd}^{M-W} and σ_{hy}^{W} are the VSE of during speech absence and speech presence. According to Eqs. (1), (4), and (5), $h_y(i)$ can be written as

$$h_y(i) = \eta(i)h_X(i) + (1-\eta(i))h_D(i), \quad \eta(i) = \frac{\xi(i)}{1+\xi(i)}, \tag{8}$$

where $h_X(i)$ is the instantaneous spectral entropy of clean-speech, and $\xi(i)$ is the instantaneous SNR $\{\xi(i) = [x(i)/d(i)]\}$. The next step is substituting Eq. (8) into Eq. (7). Since we assumed $h_X(i)$ and $h_D(i)$ are independent of each other, when the instantaneous SNR is relatively high, $\xi(i)/[1+\xi(i)] = 1$, otherwise $\xi(i)/[1+\xi(i)] \ll 1$. For low instantaneous SNR $1/[1+\xi(i)] = 1$, otherwise $1/[1+\xi(i)] \ll 1$. Therefore, $\mathrm{var}^W\{1/[1+\xi(i)]\} \approx \mathrm{var}^W\{\xi(i)/[1+\xi(i)]\} \approx 0$, then we have

$$\begin{aligned}\sigma_H = {} & \frac{M-W}{M}\sigma_{hd}^{M-W} + \frac{1}{M}\sum_{i=1}^{W}(\eta(i))^2\sigma_{hx}^W \\ & + \frac{1}{M}\sum_{i=1}^{W}(1-\eta(i))^2\sigma_{hd}^W \\ & + \frac{2W}{M}\bar{h}_X^W\bar{h}_D^W\mathrm{cov}^W(\eta(i), 1-\eta(i)) \\ & + \frac{W}{M}\left(1-\frac{W}{M}\right)\left(\bar{h}_D^{M-W} - \frac{1}{W}\sum_{i=1}^{W}(\eta(i))\bar{h}_X^W \right. \\ & \left. - \frac{1}{W}\sum_{i=1}^{W}(\eta(i))\bar{h}_D^W)\right)^2, \end{aligned} \tag{9}$$

where $\eta(i) = \xi(i)/[1+\xi(i)]$, $\bar{h}_D^W$ is the MSpE of noise during speech presences. $\mathrm{cov}^W[\eta(i), 1-\eta(i)]$ is the covariance between $\eta(i)$ and $1-\eta(i)$ over W (speech presences). Since $(1/M)\sum_1^W[\eta(i)]^2$ and $(1/M)\sum_{i=1}^W[1-\eta(i)]^2$ are the functions of global SNR, it can be seen that VSE (σ_H) depends on the global SNR.

In practice, Eq. (9) is difficult to solve because the noise corrupts clean speech; thus, σ_{hx}^W and $\bar{h}_X^W$ of clean-speech are unable to be measured directly. Instead, we assume that different clean-speech contents have a similar degree of VSE difference to a specific type of noise that share the same VSE-SNR relationship. The relationship functions, which represent the VSE-SNR relationship in different types of noise, are estimated using generated noisy-speech samples and saved as a lookup table (detailed in Sec. III B) for the SNR estimation of new noisy-speech samples. The mean values of VSE over different time samples for each SNR level are estimated for each point of the relationship function (the process is detailed in Sec. III B). This is based on the general assumption that at each SNR level, the VSE of different noisy-speech samples roughly follows a Gaussian distribution. The VSE relies on the spectral coefficients and amplitude of clean-speech and noise, whose distributions have been characterized by Gaussian models (Malah and Ephraim, 1985; Jensen *et al.*, 2005). According to Eq. (9), the distribution range relies on the variation of σ_{hd}^{M-W}, σ_{hd}^W, $\bar{h}_D^{M-W}$, $\bar{h}_D^W$, and σ_{hx}^W, $\bar{h}_X^W$ (at the same SNR), which are caused by the inherent spectral characters differences over noisy-speech samples. Their variation would cause noisy-speech samples to have different VSE at the same SNR level and possibly degrade the SNR estimation accuracy.

By assuming that the amplitude of clean-speech and noise are independent, the above variations dominated by either noise or clean-speech can be discussed separately. Specifically, σ_{hd}^{M-W}, σ_{hd}^W, are the variance of noise spectral entropy during speech absence and presence, respectively. $\bar{h}_D^{M-W}$, $\bar{h}_D^W$ are the mean of noise spectral entropy during speech absence and presence, respectively. According to Eqs. (2)–(4), the variations of σ_{hd}^{M-W}, σ_{hd}^W, $\bar{h}_D^{M-W}$, $\bar{h}_D^W$ are caused by the variance of instantaneous noise power $[\sigma_d(k,i)]$ over different noise intervals. Similarly, the variations of the variance (σ_{hx}^W) and mean $(\bar{h}_X^W)$ of clean-speech spectral entropy are caused by the variance of instantaneous clean-speech power $[\sigma_x(k,i)]$ over different clean-speech intervals. According to Eqs. (1) and (3), both of the noise and clean-speech dominated variances can be expressed by

$$\begin{cases} \sigma_d(k,i) = \sum_{j=1}^{J}\left(|g(k)(H(k,i)*d_j(i))|^2 - \frac{1}{J}\sum_{j=1}^{J}|g(k)(H(k,i)*d_j(i))|^2\right)^2, \\ \sigma_x(k,i) = \sum_{j=1}^{J}\left(|g(k)(H(k,i)*x_j(i))|^2 - \frac{1}{J}\sum_{j=1}^{J}|g(k)(H(k,i)*x_j(i))|^2\right)^2, \end{cases} \tag{10}$$

where $\bar{x}$ and $\bar{d}$ are the mean instantaneous power of clean-speech and noise and J is the total number of noisy speech intervals. If using a linear filter-bank, the terms $\sigma_d(k,i), \sigma_x(k,i)$ will be large, particularly in nonstationary noise and high SNRs. This is because the instantaneous power of both nonstationary noise and clean-speech are unstable. When $g = 1$ dB/dB, such variances will be linearly propagated to the calculated VSE and degrade the VSE-SNR relationship function.

Therefore, the main objective of the present study is to use the outputs of a nonlinear filter-bank with a compressive

https://doi.org/10.1121/10.0001168

gain to calculate the VSE in order to reduce the VSE-SNR relationship function variation.

III. METHOD

A. Reducing the VSE variation using the nonlinear pathway of the DRNL filter-bank

The nonlinear filter-bank is implemented based on the existing dual resonance nonlinear (DRNL) filter-bank model, which simulates the nonlinear response of the cochlea in the auditory system (Lopez-Poveda and Meddis, 2001). The DRNL filter-bank consists of two signal pathways. (1) The linear pathway contains a cascade of three identical first-order Gammatone band pass filters with linear gains to simulate the linear response of cochlea to stimulus at the frequencies below the centre frequency (CF). (2) The nonlinear pathway simulates the cochlear response to stimulus at or above the CFs. Each CF is characterized by a cascade of three identical first-order gammatone filters. A "broken-stick" nonlinear gain function is applied to simulate the compressive response of the cochlea. The "broken-stick" function simulates a compressed gain of 0.25 dB/dB at input sound levels above the compression threshold and applies linear gain at input sound levels below the compression threshold. The compression threshold is frequency specific. The level of the compression threshold decreases from 40 to 30 dB with increasing CFs. Processing after the "broken-stick" function comprises another three identical gammatone filters. More details of the DRNL filter-bank are provided in Lopez-Poveda and Meddis (2001).

Only the nonlinear pathway of the DNRL filter-bank is implemented in the present study. The nonlinear pathway outputs are used to calculate the VSE according to Eqs. (2)–(5). We only implement the nonlinear pathway for two reasons. First, it reduces almost half of the computational complexity. Second, the nonlinear pathway dominates the outputs of the DRNL filter-bank at relatively low sound levels (<75 dB) (Lopez-Poveda and Meddis, 2001), while in practice, and for our test purposes, the general speech level is maintained at moderate level. The implementation details of the nonlinear filter-bank are compared with that of the linear filter-bank (Liu *et al.*, 2017) in Table I. The nonlinear filter-bank is built with gammatone filters, while the linear filter-bank is built with Butterworth filters. The two types of filter differ in phase, impulse response, and frequency response shape (the slope of the filter skirt is different). Since the spectral entropy is defined as the probability associated with instantaneous power in each frequency band, we mainly consider the influence of frequency response shape difference on the VSE. To reduce the effect of the filter type difference on VSE performance, each frequency band of the two filter-banks are set to have the same bandwidth, particularly, the equivalent rectangular bandwidth (ERB) of auditory system suggested in Glasberg and Moore (1990). Although the linear pathway of the DRNL filter-bank has been removed when calculating the VSE, it has little effect on the final response of the DRNL filter-bank to inputs at low or moderate levels. As shown in Lopez-Poveda and Meddis (2001), the linear pathway only slightly increases the 10 dB (or above) cut-off frequency for input levels between 30 and 70 dB. Consequently, in the current evaluation, both filter-banks would demonstrate similar efficiency in extracting spectral information for calculating the VSE. Jürgens *et al.* (2016) also used second-order Butterworth filters with selected bandwidths to replace the gammtone filters in the DRNL filter-bank when simulating the cochlear response in their hearing model. Therefore, by setting the same ERB, the differences between Butterworth and gammtone filters are considered to be reduced and insufficient to affect VSE estimation performance. Within the linear filter-bank, each of the second-order Butterworth filters is set to have the 3 dB down bandwidth equal to the bandwidth suggested in Glasberg and Moore (1990). Since the nonlinear filter-bank contains a cascade of six gammatone filters, each filter has a bandwidth broader than the ERB suggested in Glasberg and Moore (1990) to make sure the final output of DRNL filter-bank has the bandwidth similar to that of the linear filter-bank. The bandwidth of the nonlinear pathway is calculated using the algorithm provided in Lopez-Poveda and Meddis (2001). The other parameters of the nonlinear-pathway are similar to those used in Lopez-Poveda and Meddis (2001).

In the present study, the DRNL filter-bank with

$$g = \begin{cases} a(k) & |H(k,i) * y(i)| < th(k) \\ b(k)|H(k,i) * y(i)|^{-0.75} & |H(k,i) * y(i)| \geq th(k) \end{cases}$$

TABLE I. Details of the filter banks used for the nonlinear and linear implementation. Showing values for the sample rate, filter type, filter orders, number of cascades, centre frequency of the filter, and filter bandwidth.

	Nonlinear filter-bank	Liner filter-bank
Sample rate (kHz)	16	16
Filter type	Gammatone filter	Butterworth filter
Orders	1	2
Number of cascades	6	1
Centre frequency (Hz)	250 367 540 794 1167 1714 2520 3703 5443 8000	250 367 540 794 1167 1714 2520 3703 5443 8000
Bandwidth of each single filter (Hz)	215 231 255 291 343 420 532 698 942 1300	57 71 92 122 167 232 329 470 679 985 (ERB)

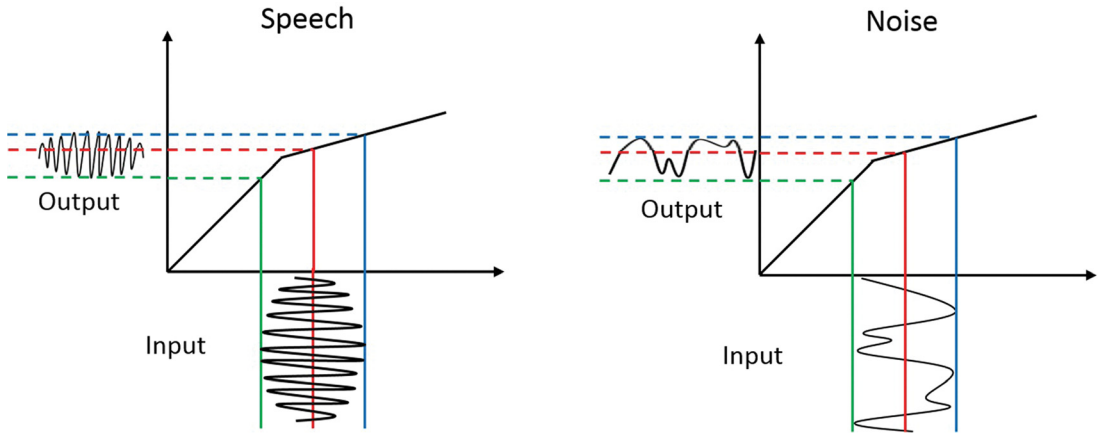

FIG. 1. (Color online) Examples of compression reducing the signal variance. The average level of the signal is marked in red. The signal below average level is marked in green. The signals above average are marked in blue. The left-hand panel shows a clean-speech signal, and the right panel shows a pure noise signal.

is used to calculate the VSE, where $th(k)$ is the compression threshold, $a(k)$ and $b(k)$ are the DRNL filter parameters (Lopez-Poveda and Meddis, 2001). Let $\bar{x}(k,i) \triangleq (1/J)\sum_{j=1}^{J}|[H(k,i)*x_j(i)]|$ and $\bar{d}(k,i) \triangleq (1/J)\sum_{j=1}^{J}|[H(k,i)*d_j(i)]|$. At the general speech level of 60 dB, we have $\bar{x}(k,i) \geq th(k)$. According to Eq. (10), the effect of compression can be discussed in the context of three different SNR conditions:

(1) For high SNRs, the $\sigma_x(k,i)$ of clean speech dominates the variation of VSE. As shown in Fig. 1 (left panel), when $|H(k,i)*x_j(i)| < \bar{x}(k,i)$ (signals between green and red lines in the left panel of Fig. 1), $\sigma_x(k,i)$ is reduced because the compressed gain reduces the $\bar{x}(k,i)$ but retains the signal $x_j(k,i)$ below $th(k)$. When $|H(k,i)*x_j(i)| > \bar{x}(k,i)$(signal between red and blue lines in the left panel of Fig. 1), $\sigma_x(k,i)$ is also reduced because the compression applies smaller gain to signals $\{|[H(k,i)*x_j(i)]|\}$ at higher levels. Note that, the overall spectrum of clean speech shows an amplitude decrease with increasing CFs (Löfqvist and Bengt, 1987). In the DRNL filter-bank, the cochlear compression over frequencies is simulated by letting $th(k)$ decrease with increasing CFs, which helps to guarantee that $\bar{x}(k,i) > th(k)$ in each frequency band.

(2) For low SNRs, the $\sigma_d(k,i)$ of noise dominates the variation of VSE, and $\bar{d}(k,i) \geq th(k)$ as shown in Fig. 1 (right panel). In both stationary and nonstationary noise, when $|H(k,i)*d_j(i)| < \bar{d}(k,i)$, if $|H(k,i)*d_j(i)| < th(k)$, then $\sigma_d(k,i)$ is reduced because the compressed gain reduces the $\bar{d}(k,i)$ but retains the signal below the $th(k)$ (signals between green and red lines in the right panel of Fig. 1). If $|H(k,i)*d_j(i)| < th(k)$, then $\sigma_d(k,i)$ is reduced because less gain is applied on $\bar{d}(k,i)$. When $|H(k,i)*d_j(i)| > \bar{d}(k,i)$, then $\sigma_d(k,i)$ is also reduced because less gain is applied to $|H(k,i)*d_j(i)|$ (as signal between red and blue lines in the right panel of Fig. 1).

(3) For moderate SNR levels, the above two cases work together to reduce $\sigma_x(k,i)$ and $\sigma_d(k,i)$. As a result, the overall variation of the noisy speech is reduced, and the variation of VSE over different noisy speech contents at each SNR is reduced.

B. SNR estimation procedure

The flow chart when using the nonlinear filter-bank to estimate the SNR is shown in Fig. 2. The noisy-speech samples are processed by the nonlinear filter-bank. The outputs of the filter-bank are used to calculate the VSE using Eqs. (2)–(5). At the same time, the noise type is detected to select a noise type specific lookup table using the noise type detection method. The lookup tables comprise the stored noise-type specific relationship functions for the SNR range between −10 and 20 dB in steps of 1 dB. The relationship functions are estimated offline using the method detailed in the following section. The calculated VSE of noisy-speech is used to compare to the values in the lookup table to find the corresponding SNR. The whole estimation process is automatic, no manual intervention or oracle information is required. The number of the relationship functions can be increased to cover more types of noise based on practical conditions of individual users frequently encountered noise background.

1. Relationship function estimation

The relationship functions are estimated for a given noise-type, making the functions noise-type specific, to

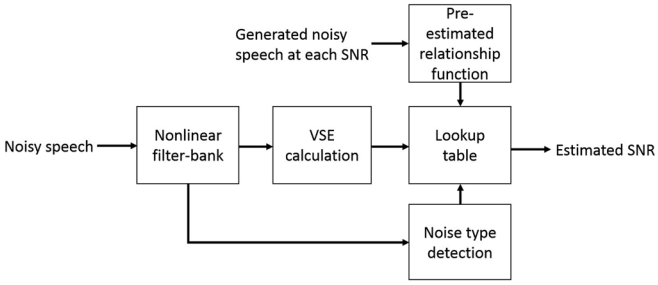

FIG. 2. Flowchart of the nonlinear filter-bank based SNR estimation.

https://doi.org/10.1121/10.0001168

reduce the SNR estimation errors over different types of noise. For each type of noise, the relationship function is estimated by calculating the mean value of VSE across different clean-speech samples corrupted by the same type of noise. The estimation consists of three stages: In stage 1, for a given SNR level, a group of random noisy-speech are generated, which are masked by the same type of added noise. In the present study, the group size is 500 random noisy-speech samples. A value of 500 noisy-speech samples was chosen because preliminary studies investigating the confidence interval of the VSE based on 1300 speech utterances (recorded from 112 speakers), found that the confidence interval of VSE stabilized at about 500. To estimate the relationship function for the different types of noise two nested procedure loops are used. The outer loop iterates over noise types and the inner loop iterates over SNRs ranging between –10 and 20 dB in 1 dB steps. The 1 dB step was regarded as a good compromise between estimation accuracy and computational efficiency. In each iteration, the noisy-speech sample is randomly generated by adding noise to clean-speech samples, following the procedure detailed in Sec. IV. Each noise sample has the same length of 1000 ms. The 1000-ms noise sample is cut from a specific type of noise resource with a random starting time. The VSE of all the random noisy speech samples is then calculated using Eqs. (2)–(5). The VSEs of all the generated noisy-speech samples are averaged to provide the estimated relationship function for a given SNR level per noise type.

2. Unknown noise type detection

In practice, the noise type is often unknown. To address unknown noise conditions, the VSE based method stores a group of pre-estimated relationship functions, which cover the noise types most often encountered in daily life. Our proposed method detects the unknown noise type and estimates the SNR using the corresponding relationship function. Unlike the method in Papadopoulos *et al.* (2016), which used a complicated noise type detection method, we classify unknown noise based on the original VSE of the noise. This is because the noise-type specific VSE-SNR relationship functions are characterized by the original VSE of the noise. At high SNRs, the relationship function mainly depends on clean speech VSE, which is independent of the change in noise type, while at low SNRs, the relationship function relies more closely on the VSE of the noise. The unknown noise type is detected by comparing the VSE of the unknown noise type with the "identification VSEs" (iVSEs) of each type of noise. Each iVSEs is estimated by calculating the mean VSEs of noise samples randomly cut from each type of noise (as described earlier). All iVSEs are saved per noise type in the detection lookup table. We use VSE to characterise noise type since different noise types differ in their signal variability. Particularly, in our case, according to Eq. (7), relationship function differences for different types of noise are mainly attributed to the original VSE of each type of noise. The VSE of a type-unknown noise sample is estimated by dividing noisy-speech into short frames (100 ms) to detect speech absences. The MSpE of each short frame can then be used to discriminate the speech absences (Shen *et al.*, 1998; Wu and Wang, 2005). Given that the MSpE of a noise-only frame is higher than that of noisy-speech frame (Wu and Wang, 2005), the frames with MSpE higher than the discrimination threshold are detected and classed as noise. In order to adapt to background noise changes, the discrimination threshold is continuously updated. If the detected frame contains speech, the threshold is updated by averaging the past thresholds; if not, the threshold is updated according to the MSpE of the noise frame. The algorithm can be expressed using the following equation:

$$P(n)=\begin{cases}0, \bar{h}(n)\le\varepsilon\rho(n-1)\\ 1, \bar{h}(n)>\varepsilon\rho(n-1),\end{cases}$$

$$\rho(n)=\begin{cases}\alpha\rho(n-1)+\dfrac{1-\alpha}{1-\delta}\bar{h}(n)-\delta\bar{h}(n-1), & \bar{h}(n)\le\varepsilon\rho(n-1)\\ \bar{h}(n), & \bar{h}(n)>\varepsilon\rho(n-1),\end{cases} \tag{11}$$

where $\bar{h}(n)$ is the MSpE of the short frame index n. $\rho(n)$ is the discrimination threshold value at the frame n, the initial value of ρ is the MSpE of the first frame. $P(n)$ is the speech absence probability. δ and α are factors used for regulating the threshold updating speed, and ε is the decision parameter. In the present study, we have $\delta=0.93$, $\alpha=0.99, \varepsilon=0.97$. After noise frame detection, the VSEs of the detected noise-frames, within each SNR estimation interval, are averaged to get $\overline{VSE}_x$. Then, the $\overline{VSE}_x$ is compared through the pre-measured iVSEs following an order from low to high. The relationship function whose iVSE is closest to $\overline{VSE}_x$ is selected for SNR estimation. For comparison, the following steps are implemented:

Step 1: Sort all the pre-measured iVSEs by their values from low to high (i.e., $iVSE_1 < iVSE_2 < iVSE_3$).
Step 2: Start with lowest identification VSE ($iVSE_1$ to compare it with the $\overline{VSE}_x$. Specifically, check If $\overline{VSE}_x \le iVSE_1 + (iVSE_{1+1} - iVSE_1/2)$; where l is the order of the iVSE in the first attempt $l=1$. If so, select the relationship function of $iVSE_1$ for SNR estimation; otherwise, go to step 3.
Step 3: Compare $\overline{VSE}_x$ with a higher iVSE ($iVSE_{l+1}$) by repeating step 2; otherwise, repeat step 3 until the If condition is met.

IV. DATASETS

The random noisy-speech samples, which are generated by adding noise samples to clean-speech samples, are used for the evaluation experiments. Next, 1300 clean-speech utterances spoken by 56 male and 56 female speakers from the AURORA (Hirsch and Pearce, 2000) resource database are divided into dataset A (500 utterances), which is used for estimating the relationship functions, and dataset B (800

utterances) which is used for evaluating SNR estimation errors. There is no overlapping between dataset A and B. Six types of talker-number specific babble noise were used for testing: 2-, 4-, 8-, 16-, 24-, and 32-talker babble derived by combining IEEE sentences (Rothauser, 1969). All sentences were normalized to have the same root-mean-square (rms) energy to generate babble noise. Although we only tested babble noise, the VSE-based method can also be used to estimate SNR in other types of noise. This can be accomplished by estimating the corresponding relationship functions and saving in the lookup tables (as described earlier). We used babble noise with different number of talkers mainly for three reasons. First, babble noise is one of the most common types of background noise encountered and therefore has high validity. It has been widely used in previous SNR estimation studies (Kim and Stern 2008; Gerkmann and Hendriks, 2012; Papadopoulos *et al*., 2016). Second, in real-life scenarios, the number of interfering talkers often varies, influencing the time-frequency characteristics of the ongoing babble noise. Cooke (2006) reported that at a given global SNR level, babble noise with a different number of talkers have different masking effects on speech perception. In addition, previous studies (Gerkmann and Hendriks, 2012; Papadopoulos *et al*., 2016) have not fully investigated the effects of babble noise comprising of different numbers of speakers on estimated SNR.

Each noise source length is 15 000 ms, and each speech utterance ranged in duration from 1000 to 3000 ms. To add a noise sample to a clean-speech sample, each noise sample (1000 ms) generated was cut from the noise resource with a random starting point (ranging from 1 to 14 000 ms). For each type of noise, the starting 10 000 ms of the resource noise was used for estimating the relationship function. The remaining duration of the resource noise was used for evaluating the results; this ensured that the training and testing noise datasets were independent. Similarly, each clean-speech sample (1000 ms) was cut from a speech resource with a random starting point (ranging from 1 to 2000 ms). The sample rate was 16 000 Hz. The noisy-speech samples were generated at SNRs ranging between −10 and 20 dB, with a steps-size of 1 dB.

V. RESULTS

To evaluate the effect of the simulated filter-bank compression to the VSE-based SNR estimation method, two experiments were presented. To begin with, as we hypothesised that the compression would increase the stability of the VSE-SNR relationship function, the coefficient of determination (known as "R^2") of the estimated relationship function derived from noisy-speech samples with a compressive filter-bank were evaluated and compared with those estimated using a linear filter-bank. R^2 (Nagelkerke, 1991) is a typical method for evaluating the goodness-of-fit of regression models. In our case, the relationship function can be considered as a data trained model. Next, we evaluated the overall effect of the compression to VSE-based SNR estimation accuracy. The SNR estimation accuracy is tested and compared with other contemporary SNR estimation methods (detailed later). The accuracy is quantified by measuring the mean absolute errors (MAE), which is widely used in literature (Narayanan and Wang, 2012; Papadopoulos *et al*., 2016). The main limitation of the MAE is that it is unable to reflect the difference between over- and under-estimation of the SNR. To address the limitations of the MAE, the under- and over-estimation of the SNR was further quantified by estimating the one-side MAE, a measure used by May *et al*., 2017).

A. Experiment 1: Coefficient of determination R^2 of the estimated relationship function

The coefficient of determination R^2 of estimated relationship functions based on a compressive (nonlinear) filter-bank (denoted as "VSE-nonlinear") to random noisy-speech samples was evaluated and compared with that of the linear-filter bank (denoted as "VSE-linear") (Liu *et al*., 2017). The relationship functions of VSE-nonlinear were estimated using speech dataset A as detailed in Sec. III. Each random noisy-speech sample was generated at a random SNR level (from −10 to 20 dB) following the procedures provided in Sec. IV. The random SNR level was generated using the "rand" function in MATLAB, which generates uniformly distributed random numbers. The speech dataset B and the testing noise dataset were used for generating random noisy-speech samples for evaluation. Two types of babble noise containing 2 and 32 talkers were used for the testing. Krishnamurthy and Hansen (2009) demonstrated that the stationary nature of babble noise decreases with decreasing number of talkers. We used 2-talker and 32-talker babble noise to test the effect of compression on highly nonstationary, and relatively stationary babble noise, respectively. Importantly, for our comparisons, both the VSE-nonlinear and VSE-linear were tested with the same noisy speech dataset.

Figure 3 plots the VSE of 200 random noisy-speech samples and the estimated VSE-SNR relationship functions of the nonlinear filter-bank (left panels) and linear filter-banks (right panels) for 2-, (upper panels) and 32-talker babble noise (lower panels). For both VSE-nonlinear and VSE-linear, the VSE of the random noise samples is concentrated more closely around the relationship function in 32-talker babble noise than in 2-talker babble noise, which is consistent with the fact that babble noise with fewer talkers is more nonstationary (Krishnamurthy and Hansen, 2009). Comparing VSE-nonlinear to VSE linear for both types of tested noise, the VSE-nonlinear demonstrates a better relationship function fit. The fit improvement is evident in 2-talker babble noise at an SNR below 5 dB. To quantify the improvement, R^2 value for each estimated relationship function was evaluated for five separate tests. In each test, all the random noisy-speech samples were re-generated to guarantee the independence of each test. The mean and standard deviation of R^2 are shown in the caption of Fig. 3. In 32 talker babble noise, the R^2 of the nonlinear filter-bank case

https://doi.org/10.1121/10.0001168

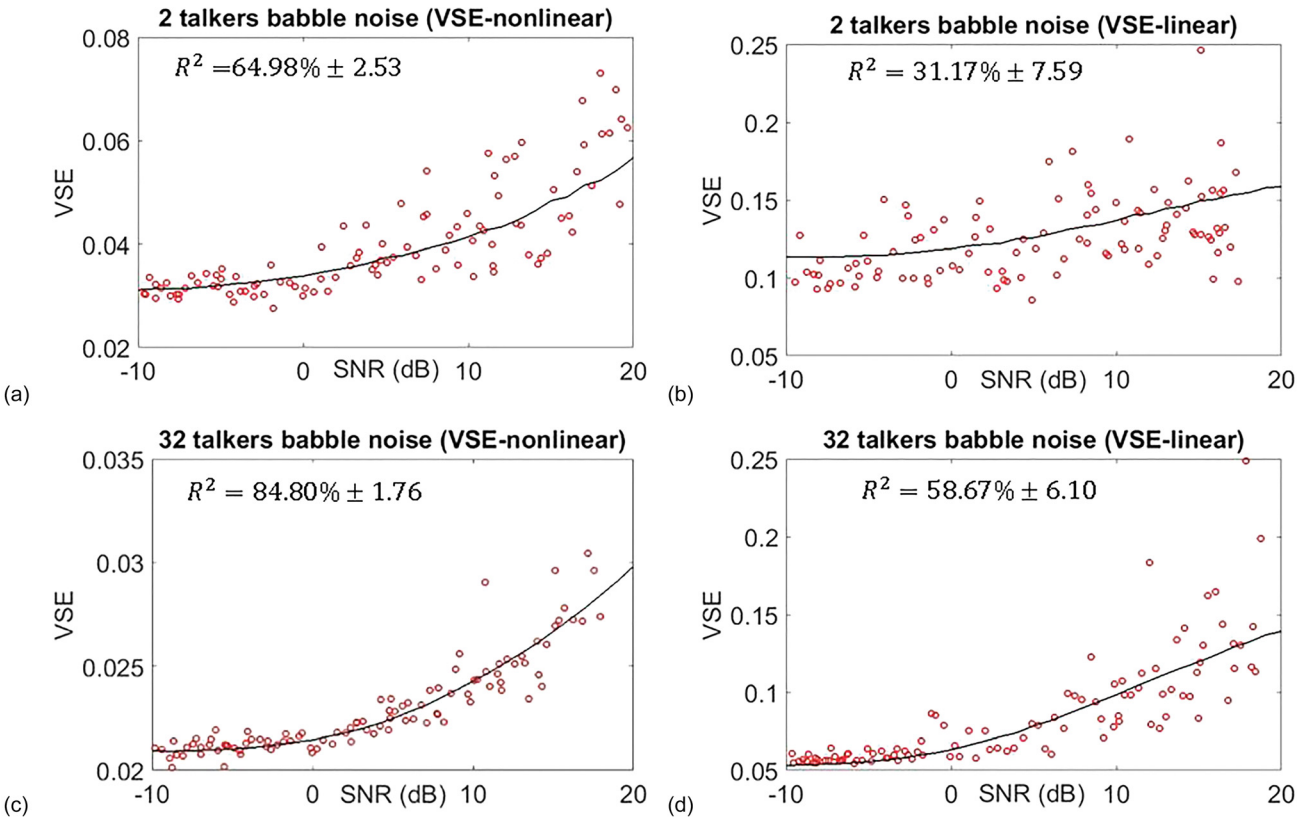

FIG. 3. (Color online) The fit of the estimated VSE-SNR relationship function to random noisy-speech samples. (a) Using a nonlinear filter-bank in 2-talker babble noise $R^2 = 64.98\% \pm 2.53$. (b) Using a linear filter-bank in 2-talker babble noise $R^2 = 31.17\% \pm 7.59$. (c) Using a nonlinear filter-bank in 32-talker babble noise. $R^2 = 84.80\% \pm 1.76$. (d) Using a linear filter-bank in 32-talker babble noise $R^2 = 58.67\% \pm 6.10$.

was about 26% (absolute percentage improvement) higher than the linear filter-bank case ($p < 0.001$). The most notable coefficient of determination improvement is shown in 2-talker babble noise where the R^2 is increased about 33% (absolute percentage improvement) ($p < 0.001$).

B. Experiment 2: SNR estimation accuracy

The SNR estimation accuracy of the VSE-nonlinear method is compared with that of the VSE-linear method, WADA method (Kim and Stern, 2008), NPE method (Gerkmann and Hendriks, 2012), and MMSE method (Erkelens *et al.*, 2007). For the VSE-nonlinear method, the SNR is estimated using noise-type specific VSE-SNR relationship functions of 2-, 4-, 8-, 16-, and 32-talker babble noise. The implementation of VSE-linear is identical to that used in Liu *et al.* (2017). The WADA method was applied using the programs provided on the website (Ellis, 2008), the default parameters were used for our purposes. The implementation of the NPE method is identical to that used in Narayanan and Wang (2012). The method of using MMSE (Erkelens *et al.*, 2007) to estimate global SNR follows the approach used in May *et al.* (2017). 2-, 4-, 8-, 16-, 24-, and 32-talker babble noise were used for testing, as explained earlier. The noise type was unseen to all the tested SNR estimation methods. The VSE-based method automatically detects the type and selects the corresponding relationship function using the noise detection method. For each type of noise, 800 noisy-speech samples are used for evaluation. Each of the noisy-speech samples is randomly generated following the procedures detailed in Sec. IV. The clean-speech data B and the testing noise dataset are used for generating these noisy-speech samples. All the test methods are evaluated with the same noisy-speech samples. Two metrics have been used to quantify the estimation accuracy. The overall estimation accuracy is quantified by measuring the MAE, which is defined by Papadopoulos *et al.* (2016),

$$MAE = \frac{1}{J}\sum_{j=1}^{J} |\xi(\mathrm{j}) - \hat{\xi}(\mathrm{j})|, \qquad \mathrm{J} = 1, \ldots, 800, \tag{12}$$

where ξ is the real SNR (the SNR used for generating the test noisy speech), and $\hat{\xi}$ is the estimated SNR. j is the index of the noisy-speech sample. In order to evaluate the differentiating between under and overestimation errors, the error measure used in May *et al.* (2017) has been applied in the present study. Both the over and under estimation errors are specified using the equations (May *et al.*, 2017)

$$one\_sided\, MAE_{over} = \frac{1}{J}\sum_{j=1}^{J} |\min(0, \xi(\mathrm{j}) - \hat{\xi}(j))|, \tag{13}$$

$$one\_sided\ \mathrm{MAE}_{under} = \frac{1}{J}\sum_{j=1}^{J} |\max(0, \xi(\mathrm{j}) - \hat{\xi}(j))|. \quad (14)$$

The averaged MAE of WADA, NPE, MMSE, linear filter-bank based VSE (VSE-linear), and the nonlinear filter-bank based VSE methods (VSE-nonlinear) across all SNR levels between −10 and 20 dB in different types of babble noise are shown in Fig. 4. The error bars represent the standard deviation of five different tests. In each test, all the random noisy-speech samples were re-generated to guarantee the independence of each test. The WADA, VSE-linear, and NPE methods show an apparent increase of MAE with decreasing talker number. In contrast, the VSE-nonlinear method is less influenced by the decrease of talker number. From 32-talker babble noise to 2-talker babble noise, the error increases by only about 1 dB, which is much lower than other methods. Moreover, the VSE-nonlinear shows the lowest MAE over most of tested noise. It only shows a MAE about 0.29 dB higher than that of MMSE in relatively stationary 32-talker babble noise. However, the improvement increases with decreasing number of talkers in babble noise. Particularly, in 2-talker babble noise, the VSE-nonlinear method shows the MAE 10 dB lower than the NPE method and about 6 dB lower than that of the WADA method. It is worth noting that the MAEs of VSE-nonlinear method are lower than 3 dB in all the tested noise types.

In comparing to the VSE-linear method, the VSE-nonlinear shows significant SNR estimation accuracy improvements in all tested babble noise. In 24- and 32-talker babble noise, the VSE-nonlinear shows the lowest improvement of about 1 dB. This indicates that the simulated compression shows less improvement compared to VSE-linear method in relatively stationary noise types, and

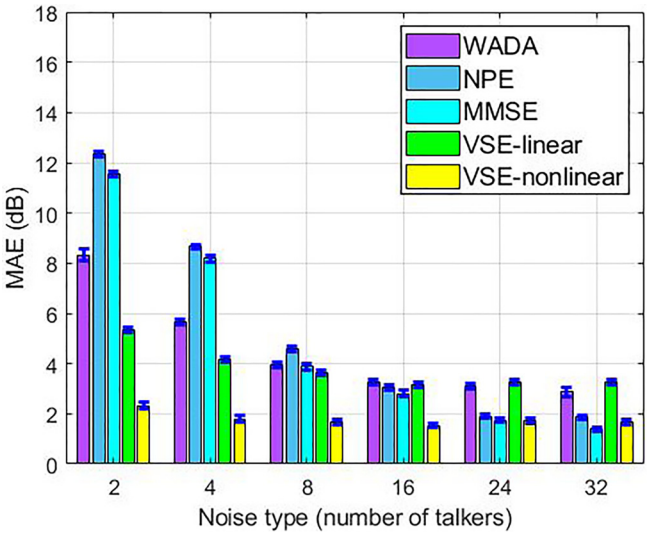

FIG. 4. (Color online) Averaged MAE over the SNR range between –10 and 20 dB in steps of 1 dB for VSE using nonlinear filter-bank, VSE using linear filter-bank, WADA, NPE, and, MMSE methods in 2, 4, 8, 16, 24, and 32-talker babble noise. The error bars represent the standard deviation for five repeat tests.

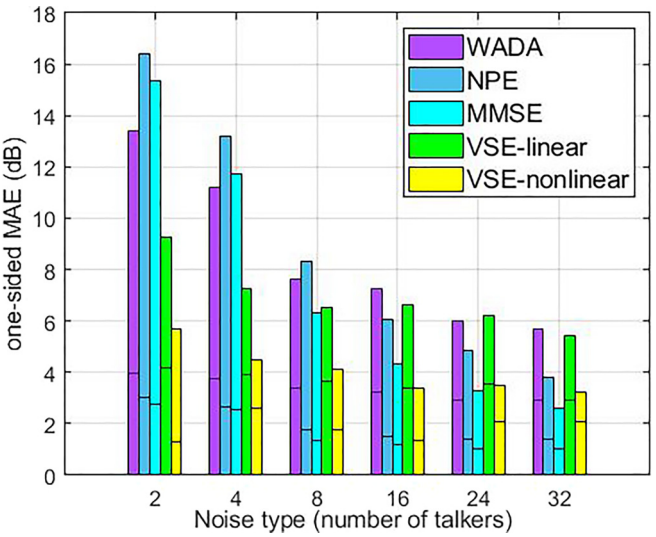

FIG. 5. (Color online) The *one_sided* MAE over the SNR range between –10 and 20 dB in steps of 1 dB for VSE using nonlinear filter-bank, VSE using linear filter-bank, WADA, NPE, and MMSE methods in 2, 4, 8, 16, 24, and 32-talker babble noise. In each column, the upper/under bars present the over/under estimation errors, respectively.

the 24- and 32-babble noises have a similar degree of stationary. The improvement increases from about 1 to about 3.5 dB when the talker number in babble noise decreases from 32 to 2. The increased improvement indicates that the VSE-nonlinear is more robust than the VSE-linear method against the highly nonstationary noise.

The *one_sided* MAE of WADA, NPE, MMSE, VSE-linear, and the VSE-nonlinear methods across SNR levels between −10 and 20 dB in different types of babble noise are shown in Fig. 5. The upper bars represent the over-estimation of the SNR, while the lower bars represent the under-estimation of the SNR. For clarity, the standard deviations of fives times tests are shown in Table II. In general, both of the over and under estimation errors of all test methods increase with the decrease of the talker number in babble noise. In the WADA, NPE, and the MMSE methods, the over-estimation errors are greater than under-estimation errors. Particularly, the MMSE method had a lower degree of over-estimation, which is consistent with the results shown in May *et al.* (2017).

TABLE II. Obtained standard deviation of the testing outputs in Fig. 5.

Number of talkers		2	4	8	16	24	32
WADA	Over	0.1029	0.111	0.1181	0.1043	0.0955	0.0801
	Under	0.1687	0.0967	0.0944	0.0971	0.0978	0.0642
NPE	Over	0.1974	0.1114	0.1162	0.1280	0.1116	0.1197
	Under	0.1302	0.1505	0.1246	0.1003	0.0606	0.0757
MMSE	Over	0.1716	0.1138	0.1198	0.1327	0.1123	0.0846
	Under	0.1104	0.1085	0.0889	0.0999	0.0813	0.0818
VSE-linear	Over	0.176	0.1344	0.1177	0.1088	0.0987	0.0715
	Under	0.12	0.0853	0.0955	0.0932	0.0985	0.0819
VSE-nonlinear	Over	0.1125	0.1243	0.0838	0.0887	0.0749	0.0671
	Under	0.1485	0.085	0.1110	0.1075	0.0791	0.0623

The VSE-nonlinear method shows further over-estimation reduction in comparison to the MMSE method. In 32-babble noise, the reduction is about 0.39 dB. As the talker number decreases, the amount of over-estimation errors reduction increases. The maximum reduction in comparing to MMSE method is shown in 2-talker babble noise, which is about 8 dB. However, in 24- and 32-talker babble noise, the under-estimation errors of the VSE-nonlinear method are greater than those of the MMSE method. When comparing the VSE-linear method, VSE-nonlinear method shows both over- and under-estimation error reduction (e.g., in 2-talker babble noise). The over-estimation error reduction is about 1 dB, while the under-estimation error reduction is about 2 dB. This indicates that the nonlinear filter-bank helps to reduce both over and under-estimation of the SNR.

VI. DISCUSSION

The present study demonstrated that the performance of the VSE based global SNR estimation method can be further improved by applying a nonlinear filter-bank, which simulates the compression response of the human cochlea. The evaluation results showed that the nonlinear filter-bank provided apparent SNR estimation improvement, particularly in less stationary noise.

The present method offers other benefits for hardware implementation. First, the proposed method is based on a filter-bank with a limited number of frequency bands, which utilizes less computational resources on devices with limited size. Second, the proposed method utilizes compression to reduce SNR estimation error.

Compared to other SNR estimation methods, the present method shows greater SNR estimation accuracy in more nonstationary noise. One of the reasons for this is that the VSE is more reliable than methods involving either tracking the noise power or estimating amplitude for estimating SNR in nonstationary noise. In Fig. 4, the WADA method, which has no tracking delays, showed SNR estimation accuracy greater than NPE and MMSE methods (both influenced by noise power tracking delays) in 4- and 2-talker babble noise. The distribution amplitude of the babble noise can be very similar to that of clean speech that degrades the SNR estimation accuracy of the WADA method. As shown in Fig. 4, the estimation errors of the WADA method significantly increase when the talker number decreases. The nonlinear filter-bank improves the performance of the VSE in estimating SNR. The performance improvement can be attributed to two aspects: (1) increases of the fit of the VSE-SNR relationship, and (2) increases of the ability on characterizing the clean speech level at low SNRs. The results demonstrate that the nonlinear function increases the fit of the VSE-SNR relationship function. However, it is not clear if the nonlinear filter-bank increases the characterizing ability of the VSE based on the present results. Future work could focus on investing how the nonlinear filter-bank would influence the ability of VSE in characterizing clean speech from the noise.

A possible limitation of the present method is that although the VSE-nonlinear method resulted in lower over-estimation errors compared to the other methods, it still has a high over-estimation error of over 5 dB in highly nonstationary noise (e.g., 2-talker babble noise). Also, most of the over-estimations occur at low SNR levels. One of the reasons for these results may be that the 2-talker babble noise has time and frequency characters more similar to that of clean speech. This similarity makes the clean speech detecting characteristics difficult to distinguished from noise characteristics, particularly when the original SNR levels are low. In consequence, the decrease in the clean speech level is relatively difficult to distinguish. Another limitation is that the VSE-nonlinear method shows greater under-estimation errors than the MMSE method in 24- and 32-talker babble noise. This might be because the VSE-nonlinear method uses less computational resources to estimate SNR that has an upper boundary of estimation accuracy, while the MMSE method performs better on tracking clean speech when noise is relatively stationary. A potential solution could be using a non-linear mapping function such as that suggested in May *et al.* (2017) to reduce the overestimations when the general SNR environmental is known and relatively stable.

VII. CONCLUSION

In summary, a nonlinear auditory filter-bank with compression was applied to calculate the VSE for SNR estimation for speech in noise. It was found that the nonlinear filter-bank improves the overall performance of a VSE-based SNR method because the compressive gain function reduces the variance of VSE-SNR relationship function. Specifically, the nonlinear filter-bank was found to improve the coefficient of determination R^2 of the estimated VSE-SNR relationship function, particularly for nonstationary noise, compared to a linear filter-bank based VSE-SNR relationship function. In 2-talker babble noise, R^2 increases by about 33% and about 26% in 32-talker babble noise. Overall, the nonlinear filter-bank based method shows over-estimation errors much lower than other compared methods. In particular, greater SNR estimation error reduction (in comparing to other methods) is shown in highly nonstationary noise (e.g, 2-talker babble noise); the over-estimation error of the present method is about 5, 9, and 8 dB lower than WADA, NPE, and MMSE methods, respectively.

Cooke, M. (**2006**). "A glimpsing model of speech perception in noise," J. Acoust. Soc. Am. **119**(3), 1562–1573.
Ellis, D. (**2008**). "Objective measures of speech quality," Columbia University https://labrosa.ee.columbia.edu/projects/snreval/ (4/26/2020).
Ephraim, Y., and Malah, D. (**1984**). "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," IEEE Speech Audio Process. **32**(6), 1109–1121.
Ephraim, Y., and Malah, D. (**1985**). "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," IEEE Speech Audio Process. **33**(2), 443–445.
Erkelens, J. S., Hendriks, R. C., Heusdens, R., and Jensen, J. (**2007**). "Minimum mean-square error estimation of discrete Fourier coefficients with generalized Gamma priors," IEEE Trans. Audio Speech Lang. Process. **15**(6), 1741–1752.

Gerkmann, T., and Hendriks, R. C. (**2012**). "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," IEEE Trans. Audio Speech Language Process. **20**(4), 1383–1393.
Ghosh, P. K., Tsiartas, A., and Narayanan, S. (**2011**). "Robust voice activity detection using Long-Term signal variability," IEEE Trans. Audio Speech Lang. Process. **19**(3), 600–613.
Glasberg, B. R., and Moore, B. C. J. (**1990**). "Derivation of auditory filter shapes from notched-noise data," Hear. Res. **47**(1–2), 103–138.
Healy, E. W., Yoho, S. E., Wang, Y., and Wang, D. (**2013**). "An algorithm to improve speech recognition in noise for hearing-impaired listeners," J. Acoust. Soc. Am. **134**(4), 3029–3038.
Hirsch, H., and Pearce, D. (**2000**). "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proceedings of ASR2000–Automatic Speech Recognition: Challenges for the New Millenium ISCA Tutorial and Research Workshop (ITRW)*, September 18–20, Paris, France.
Jensen, J., Batina, I., Hendriks, R. C., and Heusdens, R. (**2005**). "A study of the distribution of time-domain speech samples and discrete Fourier coefficients," in *Proceedings of SPS-DARTS*, April 19–20, Antwerp, Belgium, pp. 155–158.
Jürgens, T., Clark, N. R., Lecluyse, W., and Meddis, R. (**2016**). "Exploration of a physiologically-inspired hearing-aid algorithm using a computer model mimicking impaired hearing," Int. J. Audiol. **55**(6), 346–357.
Kim, C., and Stern, R. M. (**2008**). "Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis," in *Proceedings of the Ninth Annual Conference of the International Speech Communication Association*, September 22–26, Brisbane, Australia, pp. 2598–2601.
Krishnamurthy, N., and Hansen, J. H. L. (**2009**). "Babble noise: Modeling, analysis, and applications," IEEE Trans. Audio Speech Lang. Process. **17**(7), 1394–1407.
Liberman, A. M. (**1996**). *Speech: A Special Code* (MIT Press, Cambridge, MA).
Liu, F., Ifat, Y., and Demosthenous, A. (**2017**). "Variance of spectral entropy (VSE): An SNR estimator for speech enhancement in hearing aids," in *Proceedings of the International Congress on Sound and Vibration*, July 23–27, London, pp. 1–8.
Löfqvist, A., and Bengt, M. (**1987**). "Long-time average spectrum of speech and voice analysis," Folia Phon. Logopaed. **39**(5), 221–229.
Loizou, P. C., and Kim, G. (**2011**). "Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions," IEEE Trans. Audio Speech Lang. Process. **19**(1), 47–56.
Lopez-Poveda, E., and Meddis, R. (**2001**). "A human nonlinear cochlear filterbank," J. Acoust. Soc. Am. **110**(6), 3107–3118.
Martin, R., Malah, D., Cox, R. V., and Accardi, A. J. (**2004**). "A noise reduction preprocessor for mobile voice communication," EURASIP J. Appl. Signal Process. **3**, 1046–1058.
May, T., Kowalewski, B., Fereczkowski, M., and MacDonald, E. N. (**2017**). "Assessment of broadband SNR estimation for hearing aid applications," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 5–9, New Orleans, LA, pp. 231–235.
Moore, B. C. J., Peters, R. W., and Stone, M. A. (**1998**). "Benefits of linear amplification and multichannel compression for speech comprehension in backgrounds with spectral and temporal dips," J. Acoust. Soc. Am. **105**(1), 400–411.
Nagelkerke, N. J. (**1991**). "A note on a general definition of the coefficient of determination," Biometrika **78**(3), 691–692.
Narayanan, A., and Wang, D. (**2012**). "A CASA-based system for long-term SNR estimation," IEEE Trans. Audio Speech Lang. Process. **20**(9), 2518–2527.
Papadopoulos, P., Tsiartas, A., and Narayanan, S. (**2016**). "Long-term SNR estimation of speech signals in known and unknown channel conditions," IEEE/ACM Trans. Audio Speech Lang. Process. **24**(12), 2495–2506.
Plapous, C., Marro, C., and Scalart, P. (**2006**). "Improved signal-to-noise ratio estimation for speech enhancement," IEEE Trans. Audio Speech Lang. Process. **14**(6), 2098–2108.
Rothauser, E. H. (**1969**). "IEEE recommended practice for speech quality measurements," IEEE Speech Audio Process. **17**, 225–246.
Ruggero, M., Rich, N. C., Recio, A., Narayan, S. S., and Robles, L. (**1997**). "Basilar-membrane responses to tones at the base of the chinchilla cochlea," J. Acoust. Soc. Am. **101**(4), 2151–2163.
Shannon, C. E. (**1948**). "A mathematical theory of communication," Bell Syst. Tech. J. **27**(3), 379–423.
Shen, J., Hung, J., and Lee, L. (**1998**). "Robust entropy-based endpoint detection for speech recognition in noisy environments," in *Proceedings of ICSLP '98*, November 30–December 4, Sydney, Australia.
Vondrášek, M., and Pollák, P. (**2005**). "Methods for speech SNR estimation: Evaluation tool and analysis of VAD dependency," Radioengineering **14**(1), 6–11.
Wu, B. F., and Wang, K. C. (**2005**). "Robust endpoint detection algorithm based on the adaptive band-partitioning spectral entropy in adverse environments," IEEE Trans. Audio Speech Lang. Process. **13**(5), 762–774.