ACTIVE SPEECH LEVEL ESTIMATION IN NOISY SIGNALS WITH QUADRATURE NOISE SUPPRESSION

Nikolaos Dionelis and Mike Brookes

Dept. of Electrical and Electronic Engineering, Imperial College London, UK

ABSTRACT

We present a noise-robust algorithm for estimating the active level of speech, which is the average speech power during intervals of speech activity. The proposed algorithm uses the clean speech phase to remove the quadrature noise component from the short-time power spectrum of the noisy speech, as well as SNR-dependent techniques to improve the estimation. The pitch of voiced speech frames is determined using a noise-robust pitch tracker and the speech level is estimated from the energy of the pitch harmonics using the harmonic summation principle. At low noise levels, the resultant active speech level estimate is combined with that from the standardized ITU-T P.56 algorithm to give a final composite estimate. The algorithm has been evaluated using a range of noise signals and gives consistently lower errors than previous methods and than the ITU-T P.56 algorithm, which is accurate for SNR levels of above 15 dB.

Index Terms— Speech analysis; active speech level; harmonic summation

1. INTRODUCTION

The active speech level (ASL) of a speech signal is defined to be its average power during intervals when speech is present; it is equal to the total speech energy in the signal divided by the duration of speech activity. A reliable estimate of the ASL is required whenever a statistical model of speech is applied to a noisy speech signal. In this paper, we present a robust ASL estimation algorithm.

Many speech processing applications require an estimate of the ASL of a noisy speech signal. An ASL estimate is, for example, required in a speech recognizer when combining a pre-trained speech model with an adaptive noise estimate [1], [2]. In [3], an ASL estimate makes it possible to normalize the speech level prior to determining a binary mask for speech enhancement and an ASL estimate is used to determine the SNR of a noisy speech signal as an initial step in calculating a non-intrusive speech quality metric in [4].

This paper presents an algorithm for estimating the ASL that is robust to high levels of additive noise. Section 2 of the paper provides an overview of two existing algorithms: the ITU-T P.56 algorithm [6] and the composite harmonic summation (CHS) algorithm [7]. The proposed algorithm, which is a development of [7] is described in Sec. 3 and evaluated in Sec. 4. Finally, conclusions are presented in Sec. 5.

2. ANALYSIS OF THE EXISTING ALGORITHMS

2.1. The ITU-T P.56 algorithm

The ITU-T Recommendation P.56 [6] defines a standardized algorithm for objectively measuring the ASL. The algorithm first low-pass filters the rectified input signal to obtain its envelope. The speech is then defined to be active whenever the envelope has exceeded an adaptive threshold within the past 200 ms [8], [9]. The ASL is then calculated as the total energy of the signal divided by the duration of activity. The threshold used to define speech activity is circularly defined to be 15.9 dB below the calculated ASL. Throughout this paper, we define the true ASL to be the value obtained by applying the P.56 algorithm to clean speech and the SNR of a noisy signal as the ratio between the true ASL and the average noise power.

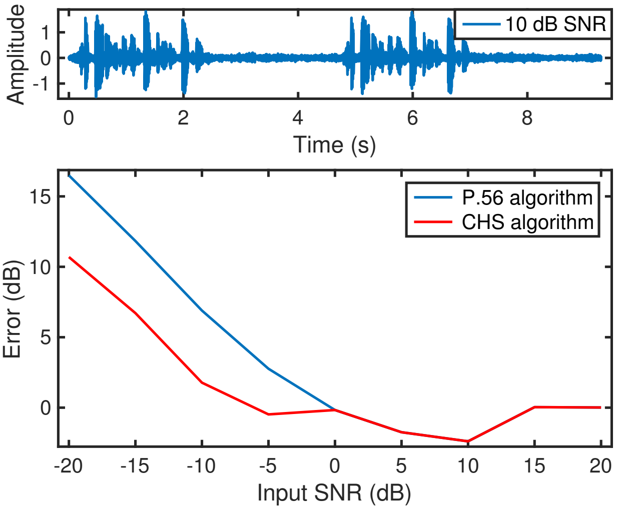

Fig. 1. The top panel shows the signals that are used in this report. In this case, babble noise at 10 dB SNR is plotted. The bottom panel shows the error for the P.56 and the composite HS (CHS) algorithms for SNR levels from -20 dB to 20 dB. For these graphs, one sentence from the TIMIT database [5] is used with babble noise.

The top panel of Fig. 1 shows a speech signal with additive babble noise at 10 dB SNR. The signal includes two segments of speech activity separated by a silent interval. The upper curve in the bottom panel shows the error in the ASL estimate of the P.56 algorithm as a function of the global input SNR of the speech signal. The ASL is calculated as the ratio of the signal energy to the active speech duration and errors arise from two opposing effects; one affecting the numerator and of this ratio and the other the denominator. First, the numerator includes the noise energy in addition to the speech energy; this results in an overestimate of the true ASL for SNR ≤ 0 dB. Second, for SNR $\leq$ 15 dB, the algorithm fails to identify the silent intervals in the speech signal and assumes that the speech is active at all times; this results in an under estimate of the true ASL for 0 dB $\leq$ SNR $\leq$ 15 dB. Figure 2 plots the activity factor (AF) of the speech in the signal from Fig. 1 versus SNR where the AF is the fraction of

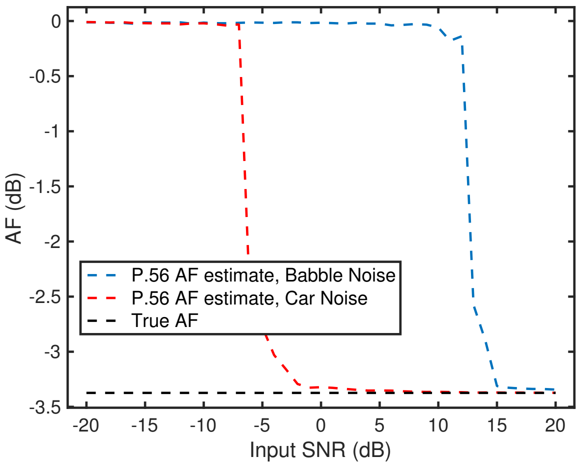

Fig. 2. The activity factor (AF) estimate of the P.56 algorithm is tested with SNR levels from -20 dB to 20 dB. One sentence from the TIMIT database [5] is used with babble noise and with car noise.

time that speech is active expressed in dB (and not as a percentage). The true AF is -3.4 dB and is shown as a horizontal dashed line (i.e. the bottom curve). The upper curve shows the estimated AF for babble noise and it can be seen that the errors are significant below 15 dB SNR. The middle curve shows the estimated AF for car noise and it can be seen that the errors are significant below -2 dB SNR. The overestimate of the AF thus varies with the noise type.

2.2. The composite harmonic summation (CHS) algorithm

An ASL estimator with improved noise robustness, the composite harmonic summation (CHS) algorithm, was presented in [7]; this uses the P.56 algorithm at high SNRs but uses an alternative approach, the harmonic summation (HS) algorithm at low SNRs. The HS algorithm uses a pitch detector [10] to identify voiced speech intervals and to track their pitch. The energy in each pitch harmonic of a voiced frame is then determined by integrating the product of the noisy speech power spectrum and a Mexican hat window centred on the frequency of the harmonic. This is illustrated in Fig. 3 which shows the power spectrum of a voiced speech frame with a larynx frequency of about 180 Hz and Mexican hat windows centred on the first six pitch harmonics. Since the Mexican hat window is symmetric and integrates to zero, a noise component in the power spectrum that is a linear function of frequency will not contribute to the integral. To determine the ASL, the CHS algorithm determines the average power in the first 15 pitch harmonics of the voiced frames and adds a constant offset of 0.85 dB to compensate for the energy of the unvoiced frames and the high harmonics. In order to reduce the errors of the ASL estimate at high SNR levels, the CHS algorithm uses the noise estimator from [11] to determine the global SNR and reverts to the P.56 estimate of ASL for SNR ≥ 4 dB. For SNR levels in the range -2 to 4 dB the CHS algorithm uses a weighted average of the P.56 and the HS estimate described above.

The errors of the CHS algorithm are shown as the lower curve in the bottom pannel of Fig. 1 where it can be seen that the error is reduced by up to 6 dB for SNR levels below -5 dB. For 0 dB $\leq$ SNR $\leq$ 15 dB, the CHS algorithm is the same as the P.56 algorithm and so it does not correct for the underestimation of the ASL in this range. This underestimation arises because the P.56 algorithm fails to detect intervals of speech inactivity at these levels of SNR.

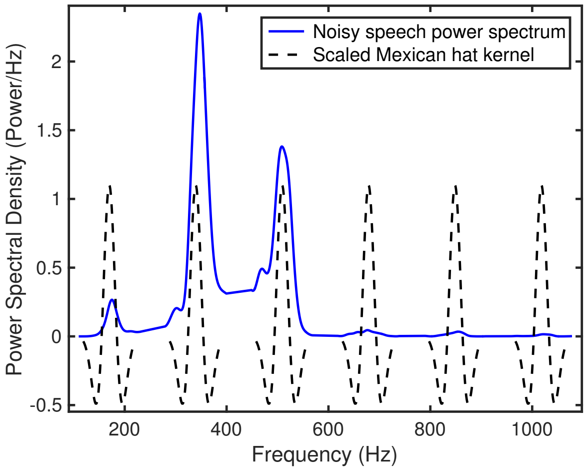

Fig. 3. The 120 Hz Mexican hat is used in the frequency domain to compute the voiced speech power. In this case, the pitch of the male speaker is approximately 180 Hz, the noise type is babble noise, the global SNR is 20 dB and the time-local SNR is higher than 20 dB.

Because the evaluation of the CHS algorithm in [7] did not include speech samples that included pauses, the degradation in performance for SNR levels in this range is not apparent from the results in [7].

3. THE PROPOSED ALGORITHM

In this section we propose an algorithm, that is based on the CHS algorithm [7] but incorporates a number of modifications that improve its performance at both low and moderate SNR levels.

The algorithm operates in the short-time Fourier transform (STFT) domain in which the complex-valued coefficients of the noisy speech in frame λ and frequency bin k are denoted as $Y(\lambda, k)$. The relation between the complex-valued DFT coefficients in the STFT domain is $Y(\lambda, k) = X(\lambda, k) + N(\lambda, k)$, where $X(\lambda, k)$ and $N(\lambda, k)$ are the clean speech and noise DFT coefficients.

3.1. Quadrature noise suppression

In the HS algorithm, the power in each pitch harmonic is estimated by integrating the product of the noisy speech power spectrum, $|Y(\lambda, k)|^2$, and a Mexican hat window centred on the harmonic. Figure 3 shows the power spectrum of a single voiced frame from the speech signal shown in Fig. 1. The pitch of this frame is approximately 180 Hz and the Mexican hat windows are shown for the first six harmonics. If the power spectrum of the noise, $|N(\lambda, k)|^2$ is a linear function of k within the support of the window, its presence will not introduce a bias into the power estimate but it will, however, increase the variance of the estimation error. If we know the phase of the clean speech coefficient, $X(\lambda, k)$, then we can reduce this variance by excluding the component of the noise that is in quadrature to the speech.

Omitting λ, k for clarity and defining $\theta_X = \angle X$, we can write

$$\mathbf{E}\left\{|Y|^2 - |X|^2\right\} = \Re(N \times e^{-j\theta_X})^2 + \Im(N \times e^{-j\theta_X})^2 \geq \Re(N \times e^{-j\theta_X})^2 = \mathbf{E}\left\{\Re(Y \times e^{-j\theta_X})^2 - |X|^2\right\} \quad (1)$$

where the expectation is over the noise phase $\angle N$ which we assume to be independent of the clean speech phase θ_X. Thus, we

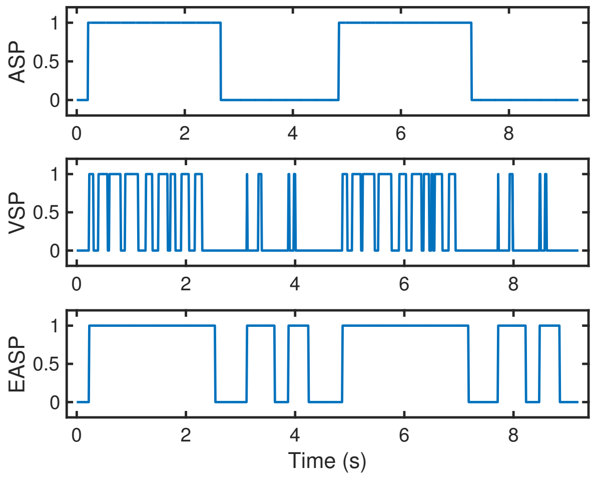

Fig. 4. The upper panel depicts the active speech presence (ASP) binary classification based on P.56 for clean speech. The centre panel shows the voiced speech presence (VSP) based on PEFAC for babble-noisy speech at 20 dB SNR. The bottom panel shows the estimated active speech presence (EASP) based on PEFAC with 250 ms hangover. Babble noise at 20 dB SNR is utilised.

see that $\widetilde{Y}^{\,2} \triangleq \Re(Y \times e^{-j\theta_X}))^2$ is a better estimate of $|X|^2$ than is $|Y|^2$. In general, we do not know $\theta_X(\lambda, k)$ a priori, so we estimate it for each λ, k using the algorithm from [12] and then use $\widetilde{Y}^{\,2}$ when estimating the power of the pitch harmonics.

We denote this variant of the HS algorithm in which quadrature noise is removed, the QHS algorithm.

3.2. SNR-dependent hangover scheme

The upper panel of Fig. 4 shows the active speech intervals, and more specifically the active speech presence (ASP) binary classification, identified by the P.56 algorithm when applied to a clean speech signal similar to that in Fig. 1. The ASP intervals are thus the true active speech intervals.

The centre panel of Fig. 4 shows the frames identified as voiced, and more specifically the voiced speech presence (VSP) binary classification, when the PEFAC algorithm [10] is applied to the speech signal with added babble noise at 20 dB SNR. The voiced frames are detected by PEFAC using a 50% probability threshold. The centre panel illustrates two types of error if the detected frames are assumed to represent speech activity: (a) within the active speech intervals some frames are unvoiced and, especially at poor SNRs, some voiced frames may be undetected, (b) some noise frames are wrongly detected as voiced. To address these issues, we apply a hangover interval to the detected voiced frames in order to estimate the intervals of speech activity. The bottom panel of Fig. 4 shows the effect of applying a hangover of 250 ms for noisy speech at 20 dB SNR. It illustrates the estimated active speech presence (EASP) binary classification based on PEFAC with the 250 ms hangover. We see that the gaps in the active speech segments have been completely filled in but that the false detections during the speech-inactive periods have also been lengthened.

The hangover extends the duration of speech activity in the denominator of the ASL ratio to compensate for the duration of the unvoiced speech frames when voiced speech frames are detected [13]. We apply an SNR-dependent hangover duration chosen to minimize the number of active/inactive classification errors. We first estimate the global SNR as the ratio between the ASL and the noise power estimated from [11]. The ASL is itself calculated as the ratio between the energy of the voiced frames estimated using the HS algorithm and the active speech duration estimated using a fixed 250 ms hangover to the voiced speech duration.

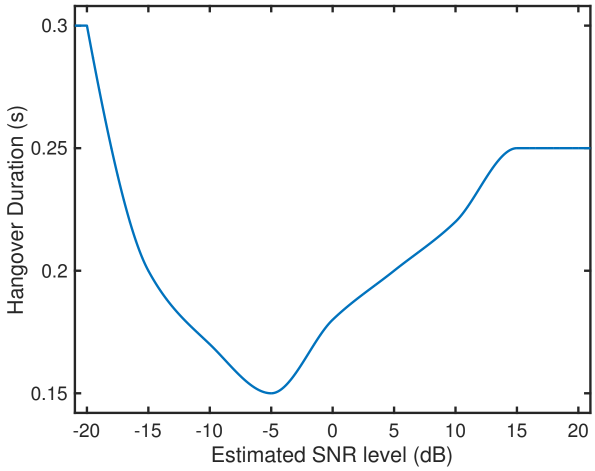

Fig. 5. The hangover duration for each SNR level is illustrated. The MSE objective function is the sum of the frame classification MSEs.

The SNR-dependent hangover duration minimizes the number of active/inactive classification errors. The MSE between the ASP and the EASP binary classification is equal to the number of classification errors. The MSE is thus between the ASP segments from the P.56 algorithm on clean speech, as in the upper panel of Fig. 4, and the voiced speech segments from the PEFAC algorithm with the hangover, as in the bottom panel of Fig. 4.

The SNR-based hangover scheme balances the tradeoff between reducing the length of the hangover due to the misclassification of certain non-voiced silence frames and increasing the length of the hangover due to the sparsity of the detected voiced frames in low SNR levels. The MSE objective function, which is the sum of the frame classification MSEs, penalises the two contradicting objectives equally and ensures that the two opposing errors do not cancel out. Due to the fact that the MSE penalises the two contradicting objectives equally, the SNR-dependent hangover does not depend heavily on the AF of the speech samples that are used for training.

Using the training dataset, defined in Sec. 4, we found the hangover duration that minimized the number of active/inactive classification errors for each value of estimated SNR in the range -20 dB $\leq$ SNR $\leq$ 20 dB. These values are plotted in Fig. 5. The optimum hangover was 250 ms at high SNRs of above 15 dB, but this decreased to 150 ms at -5 dB SNR as the number of false detections in inactive regions increased. For SNRs below -5 dB, the optimum hangover increased because at these very low SNRs the voiced frames detected within the active speech intervals become increasingly sparse.

The bottom panel of Fig. 4 shows the EASP binary classification results when 250 ms hangover is used for noisy speech at 20 dB SNR. This can be seen as an example of SNR-dependent hangover since, based on Fig. 5, for the high SNR levels of 15 dB and above, the 250 ms hangover duration minimizes the MSE objective function.

We denote this variant of the HS algorithm in which quadrature

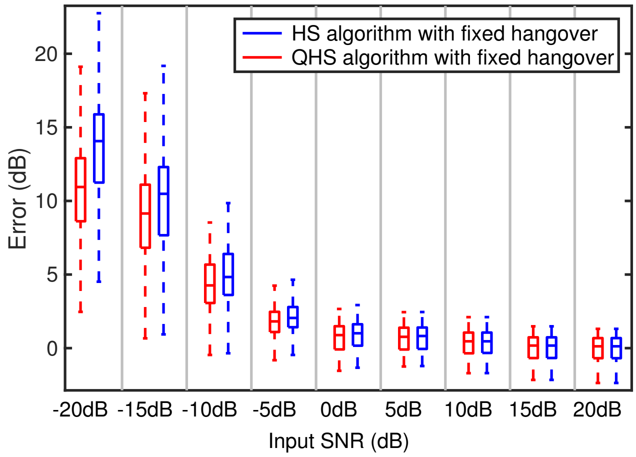

Fig. 6. The error for the HS and the QHS algorithms for SNR levels from -20 dB to 20 dB is illustrated. A fixed hangover of 250 ms is used in both algorithms. For this graph, a perfect clean phase estimator in the STFT domain is used in the QHS algorithm. A fixed offset is used to correct the ASL estimate for clean speech signals.

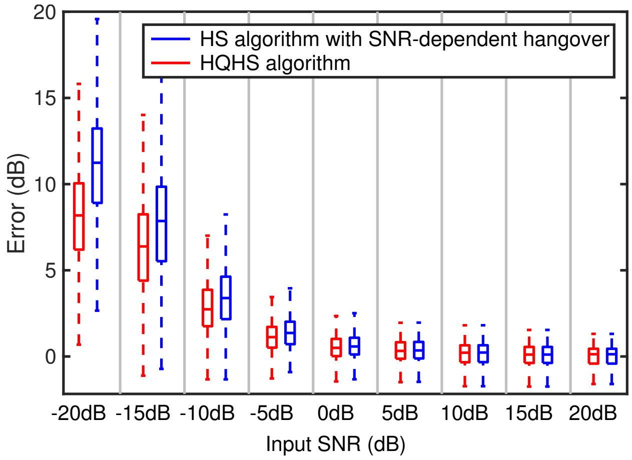

Fig. 7. The error for the HS and the HQHS algorithms for SNR levels from -20 dB to 20 dB is illustrated. A SNR-dependent hangover, based on Fig. 5, is used in both algorithms. A perfect clean phase estimator in the STFT domain is used in the HQHS algorithm. A fixed offset is used to correct the estimate for clean speech signals.

noise is removed and SNR-dependent hangover is used, the HQHS algorithm. The HQHS algorithm therefore uses QHS to estimate the harmonic energy of the voiced frames and divides this by the active speech duration determined with an SNR-dependent hangover.

3.3. SNR-dependent offset

The proposed algorithm estimates the global SNR to implement the SNR-dependent hangover scheme. This SNR estimate leads to the notion of a SNR-dependent offset rather than a fixed offset in the log-power domain. For the SNR-dependent offset, a training set has been used to obtain the offset for each SNR level. The offset is the negative value of the mean error of the ASL estimates in the training set.

We denote this variant of the HS algorithm in which quadrature noise is removed and SNR-dependent hangover and offset are used, the OHQHS algorithm.

3.4. Proposed composite algorithm

The final step of the proposed algorithm refers to the use of a composite algorithm that implements the P.56 algorithm for SNR $\geq$ 15 dB. The mean squared error is the sum of the variance and the squared bias. At poor SNR levels, the OHQHS algorithm has much lower bias than the P.56 algorithm. However, at high SNR levels of above 15 dB, the P.56 algorithm has low bias and also very low variance. Therefore, as with [7], we form a composite algorithm whose ASL estimate is a weighted average of the OHQHS and P.56 algorithms using equation (2). We will name this variant of the HS algorithm as the COHQHS algorithm. Thus, the term l_{COHQHS} denotes the estimated ASL from the COHQHS algorithm, l_{OHQHS} the estimated ASL from the OHQHS algorithm and l^{u}_{P56} the estimated ASL from the P.56 algorithm.

$$l_{COHQHS} = \rho(\gamma) \times l_{P56} + (1 - \rho(\gamma)) \times l_{OHQHS} \qquad (2)$$

The weighting factor $\rho(\gamma)$, as a function of estimated SNR γ, is determined by considering the minimisation of the error power using equation (3). The term l^u denotes the true ASL, $G(\gamma)$ is the training set and the u-superscript denotes a specific training signal.

$$\rho(\gamma) = \frac{\sum_{u \in G(\gamma)} (l^u - l^u_{OHQHS}) \times (l^u_{P56} - l^u_{OHQHS})}{\sum_{u \in G(\gamma)} (l^u_{P56} - l^u_{OHQHS})^2} \qquad (3)$$

3.5. Overview of the complete algorithm

The steps of the proposed COHQHS algorithm are as follows. The algorithm first removes the noise component that is orthogonal to speech, as described in Sec. 3.1. Quadrature noise is thus removed. Then, the algorithm uses a SNR-dependent hangover scheme to compensate for the unvoiced speech segments, as described in Sec. 3.2. Next, the algorithm uses a SNR-dependent offset as described in Sec. 3.3. Finally, the algorithm is linearly combined with the P.56 algorithm, with the SNR-dependent weighting chosen to minimize the mean squared error.

4. EVALUATION

The proposed ASL estimation algorithm is tested with signals that are created based on the clean speech files of the TIMIT database [5] and contain silence segments. For Fig. 6 to 8, 45 sentences from the TIMIT database [5] are used to create the training and test signals along with 15 noise types from the noise database in [14] for the SNR levels of -20 dB to 20 dB. Random segments of noise from the noise signals have been utilised. For training, 45 TIMIT files and 35 different speakers were used and, for testing, 45 TIMIT files and 38 different speakers were used. For each SNR level, the total number of speech and noise combinations used was 675.

The proposed algorithm uses 90 ms frames with 80 ms overlap, as in [10]. The code for the P.56 algorithm and the CHS algorithm in [15] have been used to obtain an improved CHS algorithm.

Figure 6 and Fig. 7 present the effect of the quadrature noise suppression step of the proposed algorithm. Figure 6 shows the boxplot error metrics for the corrected HS algorithm with a fixed hangover and the QHS algorithm when a perfect clean phase estimator is used.

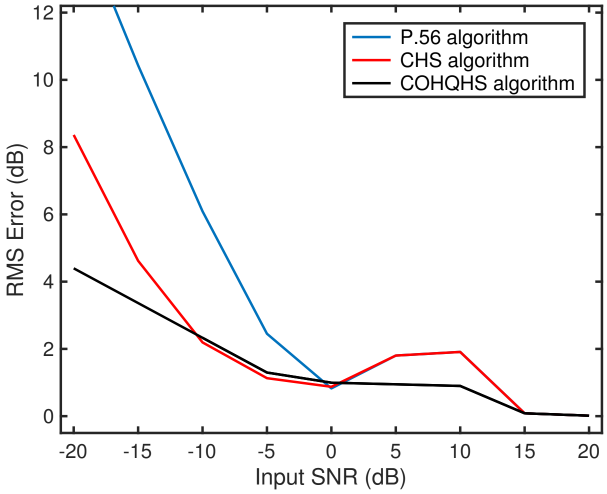

Fig. 8. The RMS error is shown for the P.56, the CHS and the proposed COHQHS algorithms for SNR levels from -20 dB to 20 dB. The COHQHS algorithm has low RMS error at $0 < \text{SNR} \leq 20$ dB.

The corrected HS algorithm refers to the HS with a fixed offset in the log-power domain. The algorithms used to generate the results in Fig. 6 used a fixed hangover of 250 ms. Both the corrected HS and the corrected QHS algorithms are more accurate than the P.56 algorithm in moderate and low SNR levels. It can be seen that the median of the error decreases for SNR ≤ 0 dB due to the quadrature noise suppression step. Figure 7 shows the relevant boxplot error metrics as Fig. 6 but with a SNR-dependent hangover scheme to compensate for the duration of the unvoiced speech frames. The difference between the two figures is the SNR-dependent hangover scheme.

Every step of the proposed ASL estimation algorithm further reduces the RMS error. Figure 6 shows that the QHS algorithm is more accurate than the HS algorithm for SNR ≤ 0 dB. Figure 6 and Fig. 7 show that the SNR-dependent hangover in the HQHS algorithm further reduces the ASL estimation error. Moreover, the OHQHS algorithm has a lower mean error than the HQHS algorithm due to the SNR-dependent offset in the log-power domain. The RMS error is the sum of the mean error squared and the error variance. The SNR-dependent offset reduces the mean error and thus the RMS error, but not the error variance of the ASL estimate.

Figure 8 shows the RMS error for the P.56, the CHS and the COHQHS algorithms. The figure illustrates that the COHQHS algorithm has lower RMS error and is consequently more accurate than the CHS algorithm for SNR levels between 0 and 15 dB. Based on Fig. 8, the COHQHS algorithm leads to an effective error reduction for a wide range of SNR values. The COHQHS algorithm has low RMS error at $0 < \text{SNR} \leq 20$ dB. However, the CHS algorithm has a slightly lower RMS error than the COHQHS algorithm for -10 dB $\leq$ SNR ≤ 0 dB due to the fact that the performance of the clean speech phase estimation algorithm from [12] deteriorates at low SNR levels and due to the fact that the fixed offset of the CHS algorithm in the log-power domain has been optimized for these SNR levels.

5. CONCLUSION

In this paper, we have presented an ASL estimation algorithm that is based on improving the HS algorithm presented in [7] in four steps. Firstly, the quadrature noise component is removed using a clean speech phase estimator. Secondly, the global SNR is estimated and a SNR-dependent hangover scheme is used to compensate for the unvoiced speech segments and, thirdly, a SNR-dependent offset in the log-power domain is utilised. The final step is the combination of the P.56 algorithm and the proposed algorithm. Using an efficient clean speech phase estimator as well as the proposed modifications of the HS algorithm, we developed the robust COHQHS algorithm. We evaluated the COHQHS algorithm and we demonstrated that it yields consistent ASL estimation accuracy improvements. Compared to the P.56 algorithm, the COHQHS algorithm is more robust to noise and, compared to the CHS algorithm, the COHQHS algorithm shows improved ASL estimation accuracy for a wide range of SNR levels.

6. REFERENCES

[1] A. Varga and R. Moore, “Hidden markov model decomposition of speech and noise,” *in Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP), pp. 845-848*, 1990.

[2] M. Gales and S. Young, “Robust continuous speech recognition using parallel model combination,” *IEEE Trans. Speech Audio Process., vol. 4, pp. 352-359*, 1996.

[3] S. Gonzalez and M. Brookes, “Mask-based enhancement for very low quality speech,” *IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, 2014.

[4] D. Kim and A. Tarraf, “ANIQUE+: A new american national standard for non-intrusive estimation of narrowband speech quality,” *Bell Labs Tech. J., vol. 12, pp. 221-236*, 2007.

[5] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallet, N. Dahlgren and V. Zue, “TIMIT accoustic-phonetic continuous speech corpus,” *Corpus LDC93S1, Linguistic Data Consortium, Philadelphia*, 1993.

[6] ITU-T, “Objective Measurement of Active Speech Level,” ITU-T Recommendation P.56, Mar. 1993.

[7] S. Gonzalez and M. Brookes, “Speech active level estimation in noisy conditions.” *International Conference in Acoustics, Speech and Signal Processing (ICASSP)*, 2013.

[8] R. W. Berry, “Speech-volume measurements on telephone circuits,” *Proc IEE, vol. 118, no. 2, pp. 335-338*, 1971.

[9] P. Kabal, “Measuring speech activity,” *MMSP Lab Technical Report, Dept. Electrical and Computer Engineering, McGill University, Tech. Rep.*, 2000.

[10] S. Gonzalez and M. Brookes, “A pitch estimation algorithm robust to high levels of noise,” *IEEE Trans. Audio, Speech, Language Processing, 22 (2): 518-530*, 2014.

[11] T. Gerkmann and R. C. Hendriks, “Unbiased MMSE-based noise power estimation with low complexity and low tracking delay,” *IEEE Trans Audio, Speech, Language Processing, 20, 1383-1393*, 2012.

[12] M. Krawczyk and T. Gerkmann, “STFT phase reconstruction in voiced speech for an improved single-channel speech enhancement,” *IEEE Trans Audio, Speech, Language Porcessing*, pp. 1931–1940, 2014.

[13] L. T. G. Hong, “Speech level estimation in acoustic noise,” Master’s thesis, Imperial College London, UK, 2014.

[14] H. Steeneken and F. Geurtsen, “Description of the RSG-10 noise database,” *TNO Institute for perception*, 1988.

[15] M. Brookes, “VOICEBOX: A speech processing toolbox for MATLAB,” 1997-2016. [Online]. Available: http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html