SPEECH ACTIVE LEVEL ESTIMATION IN NOISY CONDITIONS

Sira Gonzalez and Mike Brookes

Imperial College London
London SW7 2BT, UK

ABSTRACT

We present a new method for speech active level estimation which combines a novel algorithm based on voiced speech energy extraction with the standardized ITU-T Recommendation P.56. At poor signal-to-noise ratios, the algorithm estimates the active level by identifying intervals of voiced speech and summing the energy of the pitch harmonics in the time-frequency domain while rejecting that of the noise. We compare the performance of our method with that of ITU-T P.56 on the TIMIT database and demonstrate that it performs exceptionally well in both high and low levels of additive noise.

Index Terms— speech active level, noisy speech, fundamental frequency, pitch, speech processing.

1. INTRODUCTION

The active level of a speech signal is defined to be its average power during intervals when speech is present. The measurement of a signal's active level is an essential component in many speech processing applications. A reliable measurement of active level is needed to determine the SNR of a speech signal and in non-intrusive metrics for speech quality assessment [1]. It is also essential whenever a pre-trained speech model is combined with an estimated noise model as in the parallel model combination technique [2, 3].

The ITU-T Recommendation P.56 [4] defines a standardized method for objectively measuring the speech active level. The procedure first low-pass filters the rectified signal to obtain its envelope. The speech is then defined to be active whenever the envelope has exceeded a specified threshold within the past $200\,\text{ms}$ [5]. This threshold is circularly defined to be $15.9\,\text{dB}$ below the active level. This algorithm performs extremely well at high SNRs since the speech pauses are easily detectable in the signal envelope from their low amplitude. However, at low SNRs, the speech pauses are difficult to identify and the algorithm falsely takes some or all of the noise energy to be speech. Figure 1 shows the mean error of the ITU-T P.56 algorithm as a function of SNR for white noise. We can observe how the performance increasingly deteriorates below $5\,\text{dB}$ SNR, showing the need to develop a new speech level estimation approach based on speech characteristics that are robust to noise.

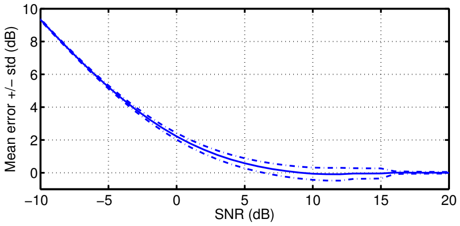


Fig. 1. Variation of P.56 mean error (solid line) plus and minus the standard deviation (dash-dot line) with SNR for white noise on 1000 utterances from the training set of the TIMIT sentence database [6].

The majority of the energy in a speech signal is concentrated in the voiced intervals. In the time-frequency domain, most of the voiced speech energy is located in a small number of harmonic peaks that remain detectable even at poor SNRs. In this paper we estimate the speech active level at low SNRs from the energy of the harmonic peaks during voiced intervals. By combining this measurement with the P.56 estimate, we obtain an algorithm that reliably estimates the active speech level even at low SNRs.

2. HARMONIC SUMMATION ALGORITHM

We assume that voiced speech can be represented as a periodic source at frequency f_0 so that our signal model in the power spectral density (PSD) domain is

$$Y(f) = \sum_{k=1}^{K} a_k \delta(f - kf_0) + N(f) \tag{1}$$

where $N(f)$ represents the power spectral density of the unwanted noise, a_k the power of the k^{th} harmonic and K is the number of harmonics. From equation (1) we note that, for this idealized signal model, all the speech energy is located at the harmonics of the fundamental frequency f_0. In practice, we process the noisy signal in overlapping frames and the energy of the harmonics is spread over a range of frequencies by the effects of the analysis window and the rate of change of

978-1-4799-0356-6/13/$31.00 ©2013 IEEE

f_0. To extract the energy of these harmonics, we need to identify the voiced speech intervals and, within these, estimate the value of f_0. This is a challenging problem at poor SNRs and a number of algorithms have been developed in recent years. In this paper we use PEFAC [7, 8], a pitch estimation algorithm robust to high levels of noise which has been shown to provide good results. We note that our proposed speech level estimation algorithm can equally be implemented using any other pitch estimator and that its robustness to noise depends heavily on the pitch estimator performance.

Once the voiced speech segments are identified and the fundamental frequency estimated, we need to measure the energy of the harmonics. For the energy of the k^{th} harmonic, we calculate a weighted integral of the frame power spectrum as $\int h(f - kf_0)Y(f)df$. The weighting function, $h(f)$, should be chosen such that:

(i) it gathers most of the harmonic energy while avoiding any interaction with adjacent harmonics,

(ii) it avoids including the energy of the noise in the harmonic energy estimate.

A weighting function that accomplish these requirements is the weighted Mexican hat wavelet, the negative normalized second derivative of a Gaussian function, which can be expressed as

$$h(f) = \left(1 - \frac{f^2}{\sigma^2}\right) e^{\frac{-f^2}{2\sigma^2}} \qquad (2)$$

To accomplish the first property, the positive part of the weighting function needs to cover the width of the harmonic and its total length needs to be restricted not to interact with adjacent harmonics. To ensure this, the support of the weighting function should lie withing $\pm \min f_0$. The width of the harmonic is mainly dependent on the window used to calculate the periodogram of the frame, as the signal frequency components, $Y_t(f)$, are convolved with the PSD of the window function, $W(f)$, to give $Z_t(f) = Y_t(f) * W(f)$. Figure 2 compares the PSD of a Hamming window having the parameters defined in Sec. 4 (dash-dot line) with the weighting function defined in (2) (solid line) with $\sigma = 15$. We can observe the fulfilment of the two requirements, as the total length is only about 100 Hz and the positive part covers the width of the harmonic.

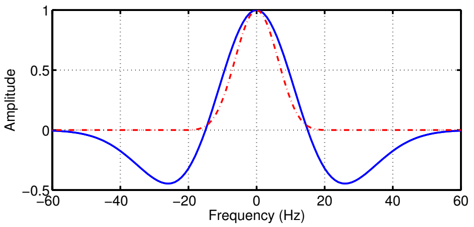


Fig. 2. Mexican hat wavelet for $\sigma = 15$ (solid line) and PSD of a Hamming window of length equal to 90 ms (dash-dot line).

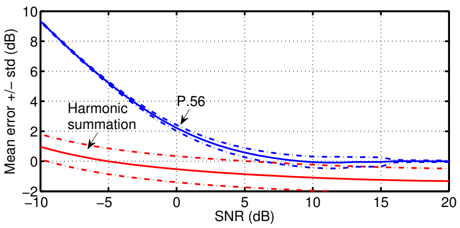


Fig. 3. Variation of the harmonic summation (red) and P.56 (blue) mean error (solid line) plus and minus the standard deviation (dash-dot line) with SNR for white noise on 1000 utterances from the training set of the TIMIT database [6].

The second requirement, the minimization of the noise contribution to the estimated harmonic energy, is accomplished since the weighting function has the property that $\int h(f)df = 0$. This means that any smoothly varying noise spectrum will be greatly attenuated.

The energy, E_t, of the first K harmonics in a voiced time frame t, is estimated as

$$E_t = \sum_{k=1}^{K} \max\left(0, \int Z_t(f)h(f - kf_0)df\right) \qquad (3)$$

The maximum function is included in (3) since the integral can be negative when the SNR is poor. The active speech level can now be estimated as

$$\hat{l}_h = \frac{1}{|V|} \sum_{t \in V} E_t \qquad (4)$$

where V represents the subset of frames which are classified as voiced by the pitch detector.

Figure 3 shows the mean and standard deviation of the estimation error as a function of SNR both for ITU-T P.56 and for the harmonic summation algorithm described above. While ITU-T P.56 obtains very good results at high SNRs, its performance degrades rapidly for negative SNRs. On the other hand, the reliability of the harmonic summation method is more constant across all SNRs but its standard deviation is higher and it understimates the speech level at high SNRs.

To compensate for the unvoiced speech energy and the underestimation of the harmonic energy we introduced an offset, β, such that

$$l_h = 10 \log_{10}\left(\hat{l}_h\right) + \beta \qquad (5)$$

The value of β is determined from a training set by minimizing the cost function $J = \sum_{u=1}^{U} \left(l^u - 10\log_{10}\left(\hat{l}_h^u\right)\right)^2$ with respect to β. This gives

$$\beta = \frac{\sum_{u=1}^{U} \left(l^u - 10\log_{10}\left(\hat{l}_h^u\right)\right)}{U} \qquad (6)$$

where l^u is the speech active level ground truth in dB for the u^{th} utterance and U is the number of utterances used for the training.

3. COMPOSITE ALGORITHM

As Fig. 3 illustrates, the P.56 active level estimate is more accurate at high SNRs but the harmonic summation method provides better results at negative SNRs. Accordingly, we combine the results from both algorithms into a new estimate that will provide reliable estimation over a larger SNR range.

In order to be able to combine the methods, we need to find a measure which identifies the transition point at which the performance of the harmonic summation method starts to be more reliable than that of ITU-T P.56. This is achieved by

$$\gamma = 10 \log_{10} \frac{\hat{l}_h}{P_N} \tag{7}$$

where $\hat{l}_h$ is defined in (4) and P_N represents the noise power estimated using the algorithm described in [9] and the implementation provided in [8]. Although it could be considered an SNR estimation, we are not aiming to estimate the SNR and consequently we are not directly concerned with the accuracy of the SNR estimate. Figure 4 shows the root mean squared error of ITU-T P.56 and the harmonic summation method for different values of γ. Three different noises were used at SNRs from -10 dB to 20 dB: white noise, car noise and babble noise. As we can observe in Fig. 4, γ provides a good way of identifying the point at which ITU-T P.56 performance starts to degrade and the harmonic summation method becomes the most reliable.

The final speech active level estimate, l_c, is calculated as a linear combination of the ITU-T P.56 estimate, l_p, and the harmonic summation method estimate, l_h,

$$l_c = \rho l_p + (1 - \rho)\, l_h \tag{8}$$

where ρ defines the contribution of each algorithm.

To determine the optimum mapping function $\rho(\gamma)$, we minimize the cost function $J = \sum_{u=1}^{U} (l - l_c)^2$ with respect to ρ and we obtain

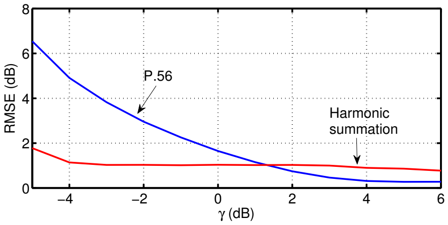


Fig. 4. Variation of the root mean squared error of P.56 and harmonic summation method with γ on 1000 utterances from the training set of the TIMIT database for white noise, car noise and babble noise.

$$\rho(\gamma) = \frac{\sum_{u \in G(\gamma)} \left(l^u - l_h^u\right)\left(l_p^u - l_h^u\right)}{\sum_{u=1}^{U} \left(l_h^u - l_p^u\right)^2} \tag{9}$$

where $G(\gamma)$ is the set of utterances having a particular value of γ.

From training data, we determined the optimal ρ for selected values of γ as shown in Table 1. We perform linear interpolation on this table for intermediate values of γ.

4. EXPERIMENTS

The test set and a subset of the training set from the TIMIT database [6] were respectively used for testing and training the algorithm. The sampling frequency of the speech material is 16 kHz. To determine the ground truth for the speech active level, ITU-T P.56 was applied on the clean speech signal.

For training and testing, noise from the RSG-10 database [10] was added to the speech files to generate the noisy signals. The calculation of SNR used ITU-T P.56 [4, 8] for the speech level and unweighted power for the noise.

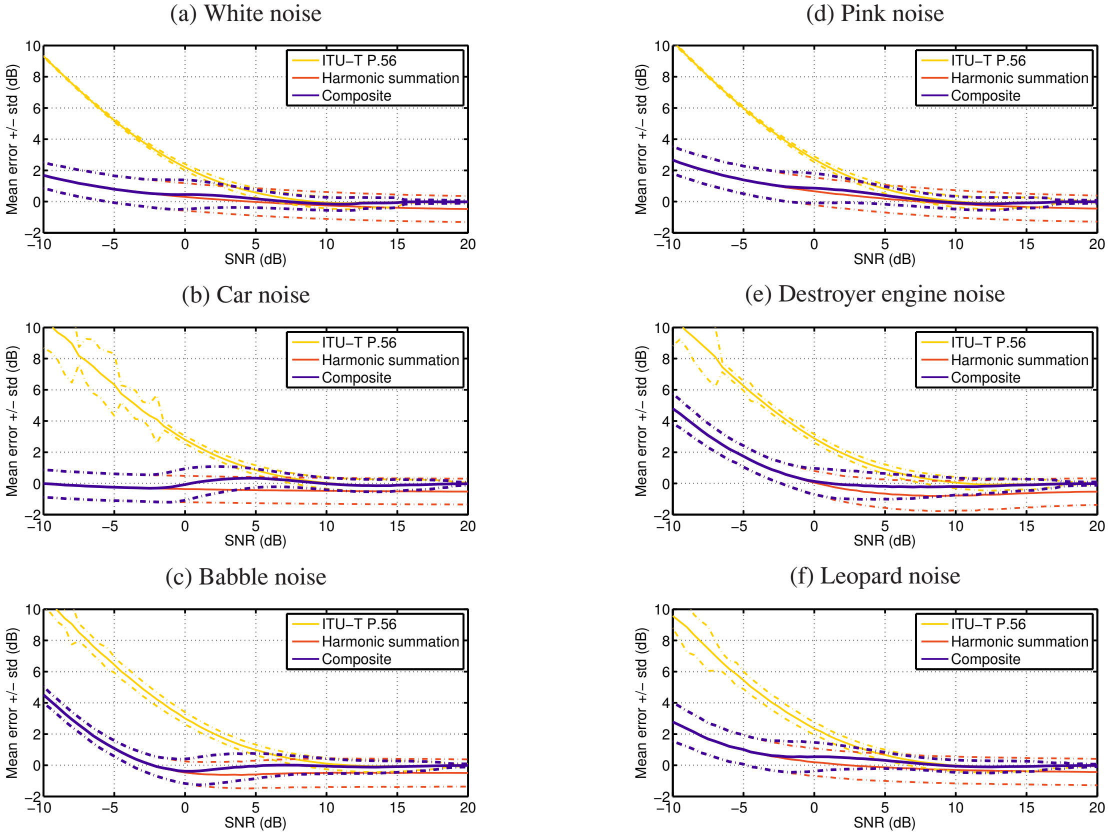
The STFT used a Hamming analysis window of 90 ms duration and the inter-frame time increment was 10 ms. This frame duration is long enough to resolve the pitch harmonics even for low values of f_0 but short enough to limit the pitch variation within a frame.

The speech active level estimation described in this paper includes a number of algorithm parameters whose values were determined empirically using the training set from the TIMIT database. The β parameter was calculated from equation (6) using 1000 utterances from the training set. Three types of noise were used at different SNRs ranging from -5 to $+5$ dB: white noise, car noise and babble noise. These three noises have different spectral characteristics and were chosen to make the results relatively independent of the noise type. The final value was set to $\beta = 0.85$.

The linear combination of ITU-T P.56 and the harmonic summation method was determined by the optimization of ρ for different values of γ. The range of γ used for the estimation was from -2 dB to 4 dB every 0.5 dB. Below $\gamma = -2$ dB, the error from the harmonic summation algorithm is much lower than that of ITU-T P.56 and $\rho = 0$ and above $\gamma = 4$ dB, the superiority of the ITU-T P.56 algorithm is clear, $\rho = 1$. Table 1 shows how, as expected, the optimum calculated value of ρ smoothly increases with γ.

Table 1. Optimized ρ values for different γ values

γ (dB)	-2	-1	0	1	2	3	4
ρ	0	0.16	0.28	0.44	0.68	0.89	1


(a) White noise

(b) Car noise

(c) Babble noise

(d) Pink noise

(e) Destroyer engine noise

(f) Leopard noise

Fig. 5. Variation of speech active level estimation accuracy on the test set of the TIMIT database with SNR for (a) white noise, (b) car noise, (c) babble noise, (d) pink noise, (e) destroyer engine noise and (f) leopard noise. The solid lines show the mean error of the estimation and the dashed lines the mean error plus/minus the standard deviation for each of the algorithms.

5. RESULTS

In this section, the performance of the proposed speech active level estimator is evaluated on the test set of the TIMIT database [6]. Six types of noise from the RSG-10 database [10] were evaluated at different SNRs from -10 to $+20$ dB: white, car, babble, pink, destroyer engine and leopard noise. While the first three kinds of noises were used in the training, the last three were new kinds of noises to the algorithm. This allows the performance evaluation of the proposed method on untrained conditions.

For each of the six noise types, Fig. 5 shows the mean and standard deviation of the estimation error for three algorithms: ITU-T P.56, the harmonic summation algorithm from Sec. 2 and the composite algorithm from Sec. 3. We observe how the combined method is able to select the best estimate at different SNRs, both on noises used for the training and on new noises. Babble and destroyer engine noise have the worst performances, with a mean error of approximately 4.5 dB at -10 dB SNR, and car noise have best performance, with a mean error close to 0 dB even at -10 dB SNR. Overall, the proposed method is able to provide a good estimation at both high and low SNRs for all the tested noise types.

6. CONCLUSIONS

In this paper we have presented a new method for estimating the speech active level which combines the ITU-T Recommendation P.56 with novel harmonic summation approach. The harmonic summation method extracts the speech harmonics' energy, providing a reliable estimation of the speech active level even at low SNRs. A fixed offset determined from training data compensates for any unvoiced speech power and for the underestimation of voiced speech power. The final speech active level estimate is calculated as a linear combination of the ITU-T P.56 estimate and the harmonic summation method estimate. The algorithm has been evaluated on the TIMIT test set with a range of noise types and extends by more than 7 dB the range of SNRs for which reliable estimation is possible.

7. REFERENCES

[1] D. S. 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.

[2] A. P. Varga and R. K. Moore, “Hidden Markov model decomposition of speech and noise,” in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 1990, pp. 845–848.

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

[4] ITU-T, *Objective Measurement of Active Speech Level*, International Telecommunications Union (ITU-T) Recommendation P.56, Mar. 1993.

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

[6] J. S. Garofolo, “Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database,” National Institute of Standards and Technology (NIST), Gaithersburg, Maryland, Technical Report, Dec. 1988.

[7] S. Gonzalez and M. Brookes, “A pitch estimation filter robust to high levels of noise (PEFAC),” in *Proc. European Signal Processing Conf. (EUSIPCO)*, Barcelona, Aug. 2011.

[8] D. M. Brookes, “VOICEBOX: A speech processing toolbox for MATLAB,” http://www.ee.imperial.ac.uk/hp/staff/dmb/voicebox/voicebox.html, 1997.

[9] T. Gerkmann and R. C. Hendriks, “Unbiased MMSE-based noise power estimation with low complexity and low tracking delay,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1383 –1393, May 2012.

[10] H. J. M. Steeneken and F. W. M. Geurtsen, “Description of the RSG.10 noise data-base,” TNO Institute for perception, Tech. Rep. IZF 1988–3, 1988.