Long-term SNR estimation using noise residuals and a two-stage deep-learning framework

Xuan Dong(✉) and Donald S. Williamson

Indiana University, Bloomington, IN 47408, USA
{xuandong,williads}@indiana.edu

Abstract. Knowing the signal-to-noise ratio of a noisy speech signal is important since it can help improve speech applications. This paper presents a two-stage approach for estimating the long-term signal-to-noise ratio (SNR) of speech signals that are corrupted by background noise. The first stage produces noise residuals from a speech separation module. The second stage then uses the residuals and a deep neural network (DNN) to predict long-term SNR. Traditional SNR estimation approaches use signal processing, unsupervised learning, or computational auditory scene analysis (CASA) techniques. We propose a deep-learning based approach, since DNNs have outperformed other techniques in several speech processing tasks. We evaluate our approach across a variety of noise types and input SNR levels, using the TIMIT speech corpus and NOISEX-92 noise database. The results show that our approach generalizes well in unseen noisy environments, and it outperforms several existing methods.

Keywords: signal-to-noise ratio estimation, speech separation, deep neural networks

1 Introduction

The signal-to-noise ratio (SNR) is a strong indicator of the amount of noise interference in a given auditory environment. Knowledge of the SNR is useful for many speech-based applications, including hearing aids [1], automatic speech recognition (ASR) [2] and speech enhancement [3], where it can be used to select model parameters or optimization strategies [4]. For a given noisy speech signal, SNR is calculated from the speech and noise components, by comparing the energy of the speech signal to the energy of the noise. Unfortunately, in real environments, the SNR must be estimated since access to the speech and noise components is not possible.

There are typically two categories for SNR estimation algorithms. The first category performs SNR estimation at the time-frequency (T-F) unit level of a signal. This is known as instantaneous or short-time SNR [5], since SNR is computed over smaller time segments. In [5], short-time SNR is computed from low-energy envelope estimates of noisy speech. In [6], a Gaussian mixture model (GMM) is used in the log-power domain to estimate the distributions of noise

and noisy speech. The decision-directed (DD) approach estimates a priori SNR with a weighted sum of the a priori SNR estimate of the prior frame and the maximum likelihood SNR estimate of the current frame [7]. The accuracy of these approaches, however, degrades when estimates are computed over long durations.

The second category performs SNR estimation at the utterance level, referred to as global or long-term SNR. The widely used NIST SNR estimation algorithm uses the bimodal observation of the short-time energy histogram of noisy speech, to infer the distributions of noise and noisy speech [8]. It then uses these distributions to calculate the peak SNR, which erroneously overestimates the true SNR. The waveform amplitude distribution analysis (WADA) approach uses a gamma distribution to model the amplitudes of clean or noisy speech using a fixed shaping parameter, and a Gaussian distribution to model the background noise [9]. WADA estimates long-term SNR by computing the maximum likelihood estimate for the shaping parameter, but WADA only performs well when the above assumptions are met, which is not always the case. Long-term SNR is also calculated from a noise power spectral density (PSD) estimator [10] or a clean speech PSD estimator [11]. A computational auditory scene analysis (CASA) based approach is proposed in [12]. The algorithm uses a ideal binary mask (IBM) to segregate noisy speech into speech dominate and noise dominate T-F regions. The energy within each region is aggregated and used to compute the long-term SNR. This unsupervised approach, however, relies on the ability of the estimated IBM to correctly label T-F units as speech or noise dominate, which does not often occur at low SNR levels. This ultimately leads to performance degradations.

The goal of our work is to improve long-term SNR estimation of noisy speech in many complex environments, since current approaches do not always perform well. Unlike prior approaches, we propose a data-driven framework that uses deep learning to perform SNR estimation. Deep neural networks (DNNs) are used, largely due to their recent success in many speech processing tasks, including automatic speech recognition and speech separation [13–15], where they have outperformed alternative approaches and been shown to generalize in unseen environments. Environmental noise plays a dominant role in degrading SNR, so our idea is to use noise distortions as an indicator of long-term SNR. Specifically, we propose a two-stage long-term SNR estimation framework. In the first stage, a speech separation system separates noisy speech into enhanced speech and noise residuals. The residuals contain mostly noise energy and can be regarded as a reasonable noise indicator for the next stage. Then the second stage uses the residuals to estimate the long-term SNR of noisy speech in a supervised manner. Our results reveal that this strategy outperforms similar single- or two-stage DNN-based approaches.

This paper is organized as follows. A detailed description of our approach is given in Section 2. Experimental results and system comparisons are given in Section 3. Section 4 concludes the discussion of the proposed system.

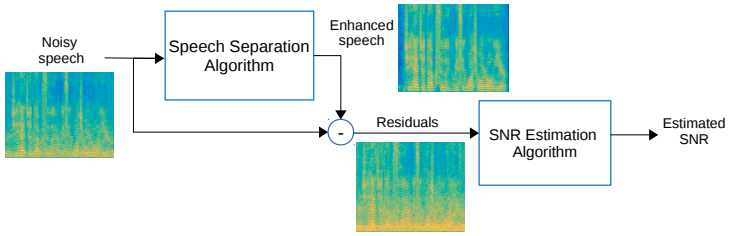

Fig. 1: The architecture of the proposed long-term SNR estimation system.

2 System Description

The proposed two-stage long-term SNR estimation approach is shown in Fig. 1. It consists of a speech separation stage and a SNR estimation stage. The goal of speech separation is to separate the target speech from background interference. We view speech separation as our first stage, but the focus of this study is to use speech separation to assist in SNR estimation. Therefore, we investigate different front-end speech separation approaches, namely, IBM estimation, ideal ratio mask (IRM) estimation, complex ideal ratio masking (cIRM), and nonnegative matrix factorization (NMF) based speech separation. Each of these stages are described below.

2.1 Speech Separation Stage

Recent approaches perform speech separation by estimating masking-based training targets [14,15]. These approaches estimate T-F masks from the noisy speech signal, and use the estimated mask to separate speech from the noise: $\hat{S}(k,f) = M(k,f) * Y(k,f)$, where $\hat{S}(k,f)$ denotes the short-time Fourier transform (STFT) of the speech estimate, $M(k,f)$ denotes the estimated T-F mask, and $Y(k,f)$ is the STFT of the noisy speech. k and f index the time and frequency dimensions, respectively. The T-F domain speech estimate is then converted to a time-domain estimate, $\hat{s}(t)$, using overlap-add synthesis. We investigate three different DNN-based T-F mask estimation approaches, namely IBM [16], IRM [17] and cIRM [18] estimation, where definitions for these three mask are shown below:

$$\begin{aligned} \text{IBM}(k,f) &= \begin{cases} 1, & \text{if } |S(k,f)| > |N(k,f)|, \\ 0, & \text{otherwise} \end{cases} \\ \text{IRM}(k,f) &= \left(\frac{|S(k,f)|^2}{|S(k,f)|^2 + |N(k,f)|^2} \right)^{0.5}, \\ \text{cIRM}(k,f) &= \frac{S(k,f)}{Y(k,f)}, \end{aligned} \tag{1}$$

$|S(k,f)|$ and $|N(k,f)|$ respectively denote the magnitude responses of the clean speech and noise. The cIRM involves complex division since $Y(k,f)$ and $S(k,f)$ are complex-valued numbers with real and imaginary components (e.g. $Y(k,f) =$

$Y_r(k,f) + jY_i(k,f)$, $S(k,f) = S_r(k,f) + jS_i(k,f)$). The IBM is a binary matrix used to label T-F units of a signal as speech or noise dominant [16], and it has been shown to improve speech intelligibility, but not perceptual speech quality. An estimated IRM often outperforms an estimated IBM [15], since it gives soft values between 0 and 1. Intuitively, the IRM represents the percentage of energy that can be attributed to speech at each T-F unit. Unlike the IBM and IRM, the cIRM enhances the magnitude and phase response of speech, since it is complex-valued. Estimated cIRMs outperform IRM-based separation when evaluated with objective metrics and human evaluations [18]. Each T-F masks impact on estimating long-term SNR, however, is not known, so we elect to separately use each of them in our front-end speech separation module.

Separate DNNs are trained to estimate each of the above mentioned T-F masks and subsequently used to perform speech separation. The structures of the DNN match those described in [15,18], where we omit details since our focus is on using speech separation to enhance long-term SNR estimation.

We alternatively use a NMF-based separation approach for our front-end speech separation stage. NMF is a model-based approach that uses trained speech and noise models (e.g. basis matrices) along with an activation matrix to separate speech from noise [19, 20]. The basis matrix represents the spectral features and the activation matrix linearly combines the spectral features to approximate a nonnegative signal. We first approximate a dictionary of clean speech signals, D, with the product of a trained basis matrix, W_{tr}, and a trained activation matrix, H_{tr} (e.g. $D \approx W_{tr}H_{tr}$). The basis and activation matrices are computed using a standard multiplicative update rule that minimizes the generalized Kullback-Leibler divergence between D and $W_{tr}H_{tr}$. To perform separation, the magnitude response of the speech estimate, $|\hat{S}(k,f)|$ is approximated as the product of W_{tr} and a new activation matrix, H_{new}, which is computed using the same multiplicative update rule and the fixed training basis matrix. Hence, $|\hat{S}| = W_{tr}H_{new}$. An estimate of the noise is computed and used along with the speech estimate to form a T-F mask. This mask is then applied to the noisy speech mixture to generate a speech estimate.

A noise residual is computed as $r(t) = y(t) - \hat{s}(t)$, where it is then provided as an input to the second stage of our approach.

2.2 Long-term SNR Estimation Stage

We train a DNN to estimate the long-term SNR of the noisy speech signal from the noise residual. A depiction of this DNN is shown in Fig. 2. Complementary features [21] are extracted from the residuals and they are provided as inputs to the DNN. These features consist of amplitude modulation spectrogram (AMS), relative spectral transform perceptual linear prediction (RASTA-PLP), and mel-frequency cepstral coefficients (MFCC). We also add delta (Δ) features to capture the temporal dynamics of the residual. We use the same parameter configuration for the complementary features as described in [18, 21], since they show success in modeling noisy speech. We tried to use log magnitude spectral

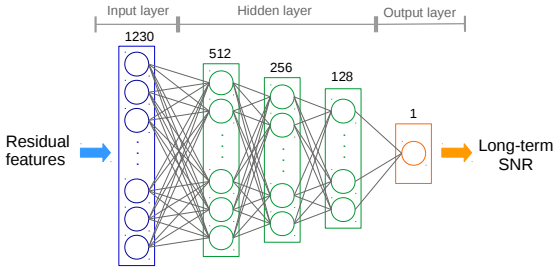

Fig. 2: Structure of the DNN that maps noise residuals to a SNR estimate.

features, Gammatone frequency (GF) features and Multi-resolution Cochleagram (MRCG) features separately as inputs, but they did not perform as well as the complementary set.

The training target is the true long-term SNR of the input noisy speech, which is calculated by the ratio of the energy of entire speech and the energy of corresponding noise, written as $\mathrm{SNR}_{\mathrm{global}} = 10\log_{10}(\mathrm{E}_{\mathrm{speech}}/\mathrm{E}_{\mathrm{noise}})$. SNR estimation, however, occurs at the time frame level, so we label each time frame with this global SNR. The DNN estimates this long-term SNR in each of the 40 ms time frames of the signal. The final estimate is generated by averaging the estimated value in each time frame. The standard back-propagation algorithm with mean-square error cost function is used for training the DNN.

The DNN has three hidden layers where each has 512, 256, and 128 units, respectively. We experimented with different number of layers and units per layer, but empirical results indicate that this structure performs best. The rectified linear (ReLU) activation function is used for the hidden units, while a linear unit is used in the output layer. After DNN training, linear regression is used to learn a linear mapping between the DNN output and the true long-term SNR. This is often done to produce better predictions for long-term SNRs that are unseen during training [12, 22].

3 Evaluation Results

3.1 Experimental Setup

All experiments are conducted with the TIMIT speech corpus [23] and NOISEX-92 noise database [24]. 600 TIMIT utterances are separately mixed with four noises: speech-shaped (SSN), cafeteria (Cafe), speech-babble (Babble) and factory floor (Factory) at 5 medium SNR levels (−6, −3, 0, 3, and 6 dB), resulting in a total of 12000 training mixtures. Random segments from the first half of each noise file is used in generating the training mixtures. These signals are used for training the DNNs of two stages, and also for generating the speech and noise models of NMF at the first stage. A separate development set is used for model selection.

Two test sets are created for evaluating the generalization performance. The first test set mixes 200 different TIMIT utterances with the same matched noise signals and at the same SNR levels as defined above. Additional mismatched SNR signals are generated at unseen low SNRs (-15, -12, and -9 dB) and unseen high SNRs (9, 12, and 15 dB), using the same matched noise signals. This results in 8800 testing mixtures. Random segments from the second half of each matched noise signal is used in generating these testing mixtures. The second test set uses 200 different clean utterances that are mixed with six unmatched noise types: cockpit, destroyer engine (Engine), machine gun (Machine), pink, tank and white noise at 11 SNR levels ranging from -15 dB to 15 dB, in 3 dB increments, producing 13200 testing mixtures.

The STFTs in the speech separation stage are computed using a Hanning window length of 40 ms, a 640 point FFT and 50% overlap between adjacent frames. Each NMF basis matrix consists of 80 basis vectors.

The accuracy of long-term SNR estimation is measured with the mean absolute error (MAE) between the true SNR t_i and estimated SNR $\hat{t}_i$ of the i-th mixture for all N testing mixtures [12].

$$\mathrm{MAE}(t, \hat{t}) = \frac{1}{N} \sum_{i=1}^{N} |(t_i - \hat{t}_i)| \tag{2}$$

3.2 Results and Discussion

In the first stage, we separately employ and compare NMF, IBM, IRM and cIRM-based speech separation approaches, and investigate their influence on SNR estimation accuracy. In addition to using the residuals that result from the above separation approaches, we separately use the true noise signal as an input to the second DNN-stage of our approach. This assumes perfect separation and we regard it as an ideal case, since it provides upper bound performance capabilities.

Table 1 shows SNR estimation results in the matched noise case, but with seen and unseen SNRs. We find that in every case the system with cIRM separation gives the best estimation especially at low SNR conditions, and its performance is close to the ideal case. This occurs because cIRM estimation outperforms the other speech separation approaches, as indicated in [18]. This reveals that improving speech separation performance can clearly improve SNR estimation accuracy. Note that the average PESQ performance is 1.81 for noisy speech, 1.88 for NMF, 1.92 for IBM, 2.23 for IRM and 2.41 for cIRM separation. Although not trained in the system, the MAE performance at high SNRs achieves the lowest average error across all approaches. This occurs because separation performance in low SNR conditions is relatively not as good as in high SNR environments. Also notice that the performance in the unseen case is approximately the same as the seen training case on average, which indicates that the proposed approach generalizes well in unseen SNR environments.

To further evaluate the generalization performance of our system, we test in matched and unmatched noise conditions. The average MAE of 11 SNR levels,

Table 1: Avg. MAE for estimating seen and unseen SNR levels of matched noise types, when applying different separation approaches.

SNR level	Ideal	NMF	IBM	IRM	cIRM
Seen Medium	1.78	4.38	4.98	3.83	**1.85**
Unseen High	1.42	2.77	2.33	1.96	**1.64**
Unseen Low	1.34	5.93	9.15	4.85	**2.01**
All	1.56	4.36	5.39	3.60	**1.86**

Table 2: Avg. MAE for SNR estimation under matched and unmatched noise conditions. The average is across all SNRs.

Noise type	NMF	IBM	IRM	cIRM
Matched	4.36	5.39	3.60	**1.86**
Unmatched	4.27	4.60	4.08	**3.91**
All	4.31	4.92	3.89	**3.09**

ranging from −15 dB to 15 dB with a 3 dB step size, is reported for each noise type, see Table 2. Not surprisingly, cIRM estimation outperforms the other approaches across matched and unmatched noise conditions.

Our approach applies linear regression to the DNN output since this can expand the SNR prediction range, which is initially limited by the range of input SNRs that are used for training. Fig. 3 shows MAE results when linear regression is and is not applied to the DNN output. Notice that the average MAE of NMF, IRM and cIRM reduce by 0.7, 0.6 and 1.1 dB, respectively, when linear regression is applied, which shows that linear mapping improves performance.

We evaluate the importance of the speech separation stage by extracting features directly from the noisy speech and then by training the SNR-estimation DNN with the noisy speech features (e.g. no separation is performed). When this is done, SNR estimation is much worse, as it does not follow the trend of input SNRs as shown in Fig. 4 (left). Alternatively, we calculate SNR directly from the speech estimate and noise residual that are produced by the speech separation stage in order to determine how important the second stage is to long-term SNR estimation. Hence, the SNR estimation DNN is not used. These estimation results are severely worse than our proposed two-stage approach, see Fig. 4 (right). This occurs because the separation stage incorrectly places some speech energy in the estimated noise signal and noise energy in the estimated speech signal. The second SNR estimation DNN helps overcome this problem. Both experiments indicate that DNN-based speech separation followed by a SNR-estimation DNN is preferred.

Furthermore, we compare our system with four state-of-the-art long-term SNR estimation methods. The first algorithm is WADA [9], which has been proven to significantly outperform NIST [8]. The second method (e.g. Noise

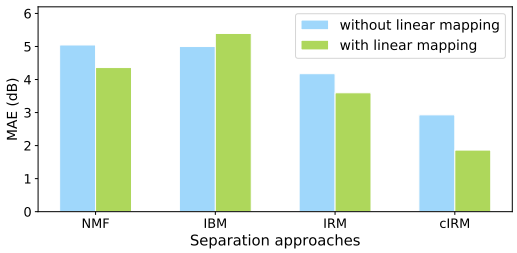

Fig. 3: Avg. MAE score when linear mapping is and is not applied.

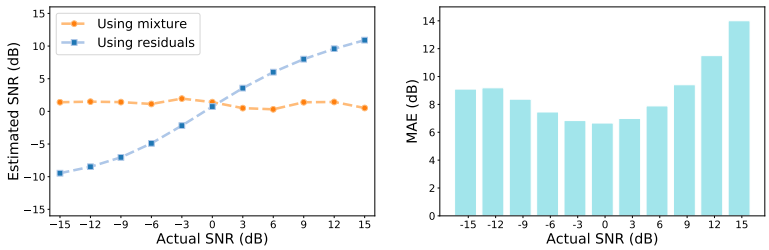

Fig. 4: Left: SNR estimation results with (e.g. residuals) and without (e.g. mixture) speech separation. Right: Avg. MAE when SNR is estimated directly from the speech and noise estimates from the speech separation stage.

PSD) estimates long-term SNR by calculating the ratio of noisy speech power and the estimated noise PSD across all time frames and frequency bins [10]. Similarly, the third algorithm (e.g. Speech PSD) uses an MMSE estimator to estimate the PSD of clean speech. The energy ratio of speech PSD and noisy speech power is used to estimate SNR [11]. The last comparison approach is a CASA-based approach that uses an estimated IBM to identify the speech-dominate and noise-dominate T-F units in a unsupervised manner [12]. The estimated IBM is then used to approximate speech and noise energies for SNR calculation. Since the proposed system with cIRM separation shows advantages over other separation approaches, it is used in the comparison and is denoted as *P*-cIRM.

As shown in Table 3, *P*-cIRM achieves the lowest MAE under matched noise conditions, and it is better by about 0.7 dB compared to the CASA approach. Compared to noise PSD and speech PSD, it is better by 4 dB and 2.8 dB, respectively. *P*-cIRM works well in unmatched noise conditions, but it is slightly outperformed by the CASA-based approach. When evaluating by SNR, *P*-cIRM shows comparative advantages over WADA, noise PSD, speech PSD, and CASA. In low SNR levels, *P*-cIRM improves by 2.5 dB compared to CASA, which also has a SNR transformation process to reduce estimation errors in low SNR conditions. *P*-cIRM also outperforms CASA at high SNRs as well. Performance for

Table 3: Comparison of the proposed system with other SNR estimation methods. * indicates that the SNR was not seen during training.

Method	Mat. Noise	Unmat. Noise	High*	Medium	Low*
WADA	8.563	10.09	8.439	6.370	13.37
Noise PSD	5.866	7.274	5.911	2.581	11.28
Speech PSD	4.737	6.513	5.274	2.349	7.530
CASA	2.599	**3.777**	2.703	**1.912**	5.170
P-cIRM	**1.864**	3.913	**2.359**	2.131	**2.596**

the CASA-based approach depends on whether it can correctly label speech and noise regions, which does not always occur at low SNRs. WADA leads to poor estimation results, since its assumption on noisy speech and noise distributions are not satisfied. Similarly, noise PSD and speech PSD assume Gaussian distributions for the noise and speech. When the background noise is non-stationary or in very low SNR levels, both noise PSD and speech PSD make relative large estimation errors, and their results are not comparable to our best performing systems.

4 Conclusion

We propose a two-stage DNN-based approach for estimating long-term SNR. The first stage generates a noise residual, and the second stage uses the residual and a DNN to predict long-term SNR. The results show that our proposed approach accurately estimates long-term SNR residuals when compared to alternative options and existing unsupervised approaches, even when tested in seen and unseen testing environments.

The results further indicate that applying better separation algorithms will obtain lower mean absolute errors. Note that our system has two independent stages. Any state-of-the-art speech separation algorithm can be used in the first stage, and more sophisticated deep learning networks can also be used in the second stage to potentially produce more accurate estimation results.

References

1. May, T., Kowalewski, B., Fereczkowski, M., MacDonald, E.: Assessment of broadband SNR estimation for hearing aid applications. In: Proc. ICASSP. (2017) 231–235
2. Ris, C., Dupont, S.: Assessing local noise level estimation methods: Application to noise robust ASR. Speech Commun. **34** (2001) 141–158
3. Cohen, I.: Relaxed statistical model for speech enhancement and a priori SNR estimation. IEEE Trans. Speech, Audio Process. **13** (2005) 870–881
4. Tchorz, J., Kollmeier, B.: SNR estimation based on amplitude modulation analysis with applications to noise suppression. IEEE Trans. Speech, Audio Process. **11** (2003) 184–192

5. Martin, R.: An efficient algorithm to estimate the instantaneous SNR of speech signals. In: Eurospeech. Volume 93. (1993) 1093–1096
6. Dat, T., Takeda, K., Itakura, F.: On-line Gaussian mixture modeling in the log-power domain for signal-to-noise ratio estimation and speech enhancement. Speech commun. **48** (2006) 1515–1527
7. Ephraim, Y., Malah, D.: Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. IEEE Trans. Audio, Speech, Lang. Process. **32** (1984) 1109–1121
8. [Online]: NIST speech signal to noise ratio measurements. Available: https://www.nist.gov/information-technology-laboratory/
9. Kim, C., Stern, R.: Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis. In: Proc. Interspeech. (2008)
10. Hendriks, R., Heusdens, R., Jensen, J.: MMSE based noise PSD tracking with low complexity. In: Proc. ICASSP. (2010) 4266–4269
11. Erkelens, J., Hendriks, R., Heusdens, R., Jensen, J.: Minimum mean-square error estimation of discrete Fourier coefficients with generalized Gamma priors. IEEE Trans. Audio, Speech, Lang. Process. **15** (2007) 1741–1752
12. Narayanan, A., Wang, D.L.: A CASA-based system for long-term SNR estimation. IEEE Trans. Audio, Speech, Lang. Process. **20** (2012) 2518–2527
13. Hinton, G., Deng, L., Yu, D., etc.: Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Sig. Process. Magazine **29** (2012) 82–97
14. Han, K., Wang, D.L.: A classification based approach to speech segregation. J. Acoust. Soc. of Amer. **132** (2012) 3475–3483
15. Wang, Y., Narayanan, A., Wang, D.L.: On training targets for supervised speech separation. IEEE Trans. Audio, Speech, Lang. Process. **22** (2014) 1849–1858
16. Wang, D.L.: On ideal binary mask as the computational goal of auditory scene analysis. Speech separation by humans and machines (2005) 181–197
17. Srinivasan, S., Roman, N., Wang, D.L.: Binary and ratio time-frequency masks for robust speech recognition. Speech Commun. **48** (2006) 1486–1501
18. Williamson, D.S., Wang, Y., Wang, D.L.: Complex ratio masking for monaural speech separation. IEEE Trans. Audio, Speech, Lang. Process. **24** (2016) 483–492
19. Virtanen, T.: Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. IEEE Trans. Audio, Speech, Lang. Process. **15** (2007) 1066–1074
20. Gemmeke, J., Virtanen, T., Hurmalainen, A.: Exemplar-based sparse representations for noise robust automatic speech recognition. IEEE Trans. Audio, Speech, Lang. Process. **19** (2011) 2067–2080
21. Wang, Y., Han, K., Wang, D.L.: Exploring monaural features for classification-based speech segregation. IEEE Trans. Audio, Speech, Lang. Process. **21** (2013) 270–279
22. Papadopoulos, P., Tsiartas, A., Narayanan, S.: Long-term SNR estimation of speech signals in known and unknown channel conditions. IEEE Trans. Audio, Speech, Lang. Process. **24** (2016) 2495–2506
23. Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallett, D.: DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1. NASA tech. report **93** (1993)
24. Varga, A., M.Steeneken, H.J.: Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems. Speech Commun. **12**(3) (1993) 247 – 251