MASK-BASED ENHANCEMENT FOR VERY LOW QUALITY SPEECH

Sira Gonzalez and Mike Brookes

Imperial College London
London SW7 2AZ, UK

ABSTRACT

We propose a mask-based enhancer for very low quality speech that is able to preserve important cues in a noise-robust manner by identifying the time-frequency regions that contain significant speech energy. We use a classifier to estimate a time-frequency mask from an input feature set that provides information about the energy distribution of both voiced and unvoiced speech. We evaluate the enhancer on a range of noisy speech signals and demonstrate that it yields consistent improvements in an objective intelligibility measure.

Index Terms— Speech enhancement, binary mask, speech intelligibility

1. INTRODUCTION

The perceived quality, and in more severe cases the intelligibility, of speech signals is impaired by the adverse noise conditions that can be encountered in real environments. Many of the numerous approaches to single-channel speech enhancement process the signal in a transform domain in which both speech and noise signals are sparse. In the popular techniques based on spectral subtraction [1] or minimum mean square error (MMSE) [2, 3], the speech is enhanced by applying an adaptive gain function in the time-frequency domain. Although these approaches normally aim to estimate the clean speech by applying a continuous gain, the goal of the more recently proposed time-frequency binary mask approaches is to retain speech information by using binary gain values. The principal advantage of the binary mask approach over other state-of-the-art algorithms operating in the time-frequency domain is that the problem of enhancement is changed from one of gain estimation to one of classification.

A speech enhancer using binary masks was introduced in [4, 5]. The classification of each time-frequency cell was based on the likelihood ratio of two Gaussian mixture models (GMMs) trained on time-frequency cells whose local whose local signal-to-noise ratio (SNR) was respectively above or below a threshold. The enhancer was evaluated on noisy speech at -5 and $0\,$dB and consistently improved the subjective intelligibility. Binary masks for speech enhancement have also been estimated using support vector machines (SVMs) [6, 7], deep neural networks [8, 9] and sparse coding techniques [10].

Our aim in this paper is to estimate the location of the time-frequency regions whose speech energy is above a frequency-dependent threshold. We extract features from the speech by using robust algorithms for detecting voiced speech, identifying its pitch, estimating the speech active level and localizing sibilant phones. In this paper, we focus on the estimation of the mask by exploiting the information captured by these algorithms.

2. GOAL OF MASK ESTIMATION

Early mask estimation algorithms used as their target the "ideal binary mask" (IBM) [11] obtained with oracle knowledge of the clean speech signal by comparing the SNR in each time-frequency bin to a fixed threshold. Many studies [12, 13] have shown that applying an IBM to noisy speech can provide perfect intelligibility for a range of fixed thresholds. The "target binary mask" (TBM) [14] achieves the same intelligibility as the IBM but eliminates dependency on the SNR by comparing the clean speech short-time spectra to the long-term average speech spectrum (LTASS) of the speaker. It has been found in [15] that the LTASS of speech signals is largely independent of both speaker and language and can therefore be represented by a universal LTASS. Accordingly, we here propose as the goal of our mask estimator, a speaker-independent universal TBM (UTBM) which is defined by

$$\mathrm{UTBM}(t,f) = \begin{cases} 1 & \text{if } S(t,f) > L(f) + \alpha + \mathrm{LC}, \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

where all quantities are in dB. $S(t,f)$ is the speech power at each time-frequency bin, α is the active level [16] of the input speech, LC is a fixed threshold called the "local criterion" and $L(f)$ is the universal LTASS spectrum from [17] normalized to an active level of 0 dB.

We evaluate the predicted intelligibility of the UTBM by using the STOI measure [19], which has been shown to give accurate predictions of the intelligibility of speech that has been enhanced by time-frequency gain modification. A comparison between the predicted intelligibility versus LC for

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

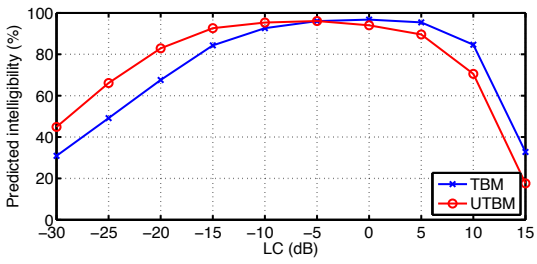


Fig. 1. Average predicted intelligibility using STOI over 98 speech segments of 5 s duration from 4 speakers from the SAM database [18]. The calculated TBM and UTBM for different LC values have been used to modulate speech shape noise.

TBM and UTBM is provided in Fig. 1. The TBM and UTBM were calculated for different LC values and used to modulate speech shaped noise. We can observe that both masks follow a similar intelligibility pattern but with a horizontal shift of 5 dB. The center of the high intelligibility range is at LC $= -5$ dB for UTBM and at LC $= 0$ dB for TBM. In general, speech will be intelligible irrespective of background noise as long as its high energy features are preserved.

3. SYSTEM OVERVIEW

A block diagram of the binary mask estimation system is shown in Fig. 2, which illustrates the steps of the mask estimation. The purpose of the estimation system is to determine a mask value, $M(t, f_e)$, for each time frame, t, and each frequency bin, f_e, that approximates the UTBM target, $\hat{M}(t, f_e)$. In the training step, the inputs to the classifier training block for each time frame consists of a set of 145 features derived from the noisy training signal, $y(\tau)$, together with the corresponding binary-valued mask target, $M(t, f_e)$, derived from the clean speech, $s(\tau)$. In the mask estimation phase, the mask, $\hat{M}(t, f_e)$, is estimated from the 145 input features alone.

3.1. Feature estimation

The selected feature set aims to provide information about the energy distribution of the speech. The feature set, explained below in detail, captures information about the presence of voiced speech and its fundamental frequency and also about the presence of sibilant speech. Moreover, the feature set also includes the normalized noisy speech and a noise estimate, which provides information about the SNR at each time-frequency bin.

In the next subsections, we explain the various processing blocks in Fig. 2 which are used to extract the features.

3.1.1. Level normalization

To ensure that classification is independent of the signal input level, the first step of the system is the power normalization of the speech component of the noisy speech signal, $y(\tau)$. The speech active level is estimated using the noise-robust algorithm described in [20] and the normalization is performed such that:

$$\overline{y}(\tau) = 10^{-l_c/20} y(\tau) \quad (2)$$

where l_c is the estimated active speech level in dB and $\overline{y}(\tau)$ the normalized signal. In the experimental results presented below, the power normalization is performed over the entire duration of the utterance.

3.1.2. Pitch and voiced speech estimator

Most voiced speech energy is concentrated at the fundamental frequency and its harmonics. Therefore, identifying voiced speech segments and estimating their fundamental frequency makes it possible to locate high speech energy regions. For each time-frame, t, the PEFAC algorithm [21, 22] provides to the classifier a fundamental frequency estimate, $\hat{f}_0(t)$, and a voiced-speech probability, $p_v(t)$.

3.1.3. Sibilant speech detector

Identifying time-frames which contain sibilant phones is important for the preservation of aperiodic speech energy at high frequencies. Furthermore, an estimation of the power spectrum of the sibilant phone helps to identify the frequency bands containing most of the sibilant speech energy. The algorithm in [23], which locates sibilant phones, is used to extract for each time-frame t the sibilant speech probability, $p_s(t)$, and a 14-component vector containing the normalized sibilant power spectrum estimate in 500 Hz bands from 1.5 kHz to 8 kHz, $\bar{b}(t, f_l)$.

3.1.4. Time-frequency decomposition

The inclusion of the normalized noisy speech periodogram and the noise estimation as parameters aids the mask estimation algorithm by providing information about the energy distribution across frequency of both speech and noise. The normalized input signal, $\overline{y}(\tau)$, is transformed into the time-frequency domain using the short-time Fourier transform (STFT).

The spectrum of each frame is interpolated onto 64 ERB spaced frequency bands ranging from 40 Hz to 8 kHz. The use of the ERB frequency scale [24] ensures that the frequency bands correspond to the spectral resolution of the ear. The output of the time-frequency transformation, $\overline{Y}(t, f_e)$, is used as a parameter for the classifier together with a noise estimate, $\hat{N}(t, f_e)$. The noise periodogram is estimated using the algorithm described in [25] and the implementation provided in [22].

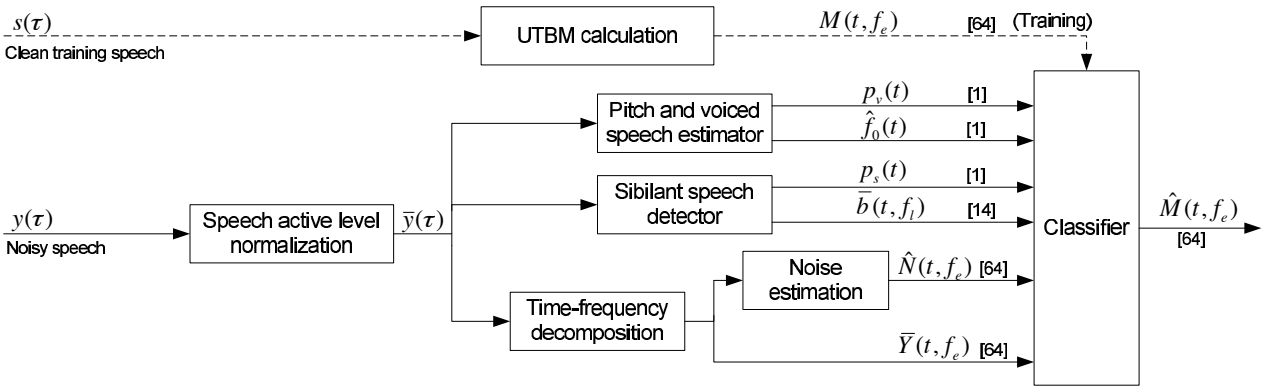


Fig. 2. Block diagram of the mask estimation system proposed. Signal vector dimensions are indicated in brackets.

3.2. Classifier

A non-parametric classification and regression tree (CART) [26] is used to generate the mask. The CART approach is convenient to handle the heterogeneous nature of the input parameters and the complex relationship between them and the target mask. The CART is trained as a regression tree using the binary valued UTBM mask values as the target. The classifier outputs lie in the range 0 to 1 and can be interpreted as the probability that the corresponding time-frequency bin energy lies above the UTBM energy threshold. The estimated probability can than be converted to a binary value by comparing it with a threshold.

4. EXPERIMENTS

The training set and the test set from the TIMIT database [27], which are from distinct speakers, were respectively used for training and testing the algorithm. The sampling frequency of the speech material is 16 kHz. To determine the ground truth for the binary mask, the UTBM was calculated for each utterance from the clean speech signal. The LC parameter was set to -5 dB, which, as was shown in Fig. 1, provides the best intelligibility results.

The STFT used a Hamming analysis window of 90 ms duration and an inter-frame time increment of 22.5 ms. The length of the window was chosen so that speech harmonics could be resolved for all f_0 values. The frame overlap of 75% results in perfect reconstruction with the Hamming window used for both analysis and synthesis. For mask estimation, the spectrum of each frame was interpolated onto 64 ERB spaced frequency bands ranging from 40 Hz to 8 kHz. We expect this frequency resolution to provide good performance since high intelligibility was found in [28] when using more than 16 frequency bands.

The regression tree was trained using the MATLAB implementation from the statistics toolbox. We used 300 TIMIT utterances from the training set mixed with 12 noises from the RSG-10 database [29]. The noise types included: factory, babble, buccaneer and F16 fighter jets, engine room, operation room, HF radio channel, leopard and M109 tank, pink, car and white at SNRs between -5 dB to $+9$ dB in 2 dB steps. The calculation of the SNR used ITU-T P.56 [16, 22] for the speech level and unweighted power for the noise. A separate regression tree was trained for each of the 64 frequency bands. The input to each regression tree contained the entire feature vector, rather than just its local frequency components.

5. RESULTS

The performance of the mask estimation was evaluated using 100 utterances from the test set of the TIMIT database mixed with noises from both the RSG-10 database [29] and the ITU-T P.501 standard [30]. SNRs from -5 to $+10$ dB were used for evaluation. For each trial, a random segment within the noise file was selected, so that the actual noise samples were distinct.

Performance evaluation used the intrusive objective intelligibility measure STOI [19]. This objective algorithm provides a value between 0 and 1 which has been found to have a monotonic relation with the subjective speech-intelligibility [19].

5.1. Continuous versus binary-valued masks

We first evaluate the performance of the continuous versus the binary gain mask. To define the binary mask, we set a probability threshold, p_b, above which the mask is set to 1

$$\hat{M}_B(t,f) = \begin{cases} 1 & \text{if } \hat{M}(t,f) > p_b, \\ 0 & \text{otherwise.} \end{cases} \quad (3)$$

where $\hat{M}_B(t,f)$ and $\hat{M}(t,f)$ represent the binary and continuous gain mask respectively. We evaluated the results for different p_b on 100 utterances from the test set on the same noise types used for training. It was found that the highest STOI values were achieved when using the continuous gain mask which we therefore use in the evaluations below. An example for factory noise at -5 dB SNR is shown in Fig. (3),

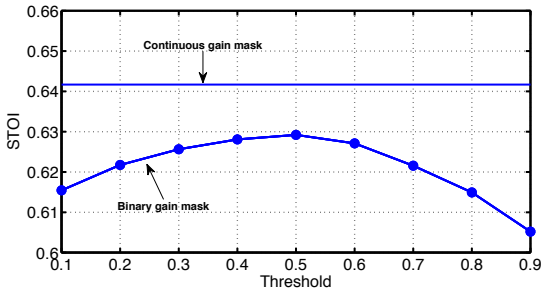


Fig. 3. STOI values for the continuous gain mask and the different binary masks for factory noise at -5 dB SNR. The STOI values are the average over 100 utterances.

where the STOI value for the continuous gain mask outperforms the binary mask estimated for any tested threshold.

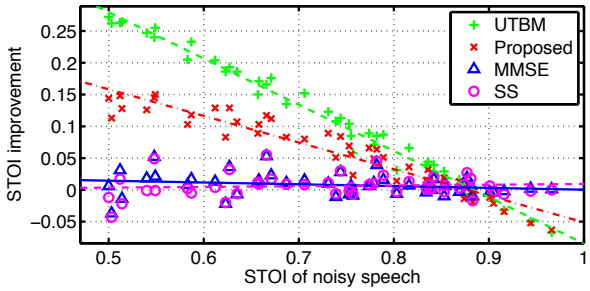
5.2. Evaluation on seen and unseen noise types

For performance comparison, the log-MMSE algorithm [3, 22], and the spectral subtraction [1, 22] speech enhancement algorithm were included in the evaluation. In both cases, the noise was estimated using the algorithm described in [25, 22], the same one used for the proposed mask estimation.

The results obtained for the three evaluated algorithms and the oracle mask are shown in Fig. 4, where the STOI improvement is plotted versus the STOI of the noisy speech. Fig. 4(a) shows the results for all 12 seen noise types used for the training and Fig. 4(b) for 10 noise types of the ITU-T P.501 standard [30] and 3 noise types from RSG-10 database [29] which were not used for training. Each data point gives the average STOI value and average STOI improvement over 100 test utterances using a specific noise type at a specific SNR. The straight lines represent the least-squares linear fit to the data points for each speech enhancement method. On average, in both seen and unseen conditions, the MMSE (blue triangles, solid blue line) and the spectral subtraction (pink circles, pink dashed line) algorithms do not change the input STOI value substantially. This is consistent with the results in [31], where no speech enhancement system was found to improve intelligibility significantly.

However, in Fig. 4(a) we can observe how the proposed mask is consistently able to improve the STOI of noisy speech for values below 0.8. It is worth noting that when the STOI value is above 0.7, the speech intelligibility is very high [19] and the impact on intelligibility of small changes to the STOI score will be insignificant. In particular, for noisy speech STOI values above 0.9, the small decreases in STOI introduced by our proposed algorithm will not affect intelligibility. The oracle UTBM mask has similar performance to the estimated mask for high STOI values while providing a STOI improvement of approximately 0.25 for an initial STOI of 0.5 versus the 0.15 improvement of the estimated mask. The

(a) Seen noise types



(b) Unseen noise types

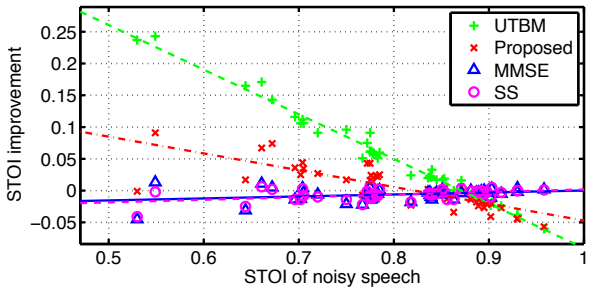


Fig. 4. STOI improvement using the proposed algorithm versus the STOI of the noisy signal for (a) seen noise types and (b) unseen noise types. The STOI values are the average over 100 utterances. The straight lines in the figure are least-squares linear fits to the data points.

results on unseen noise types are shown in Fig. 4(b). Due to the limited number of noises used for the training, our algorithm does not generalize well on all types of unseen noise and the results (indicated by red x) are not as consistent as for the seen noises. However, the proposed algorithm trend, indicated by the linear fit to the data (red dash-dot line), is to increase the STOI value when the input STOI is low, although the average increase is about half that obtained on seen noise types. We see that the oracle mask results (indicated by green +) are very similar to those shown in Fig. 4(a) for the seen noise types.

6. CONCLUSIONS

In this paper we have presented a mask-based algorithm that is able to increase the predicted intelligibility calculated using the objective STOI measure. We extract 145 features per frame from the noisy speech using robust algorithms and train a regression tree for each frequency band using the UTBM as a target. The proposed mask estimation algorithm was evaluated on the TIMIT test set with a variety of noise types. We conclude that the proposed algorithm is able to increase the predicted intelligibility for noise types seen in the training while maintaining or increasing the predicted intelligibility on unseen noise types.

7. REFERENCES

[1] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 2, pp. 113–120, Apr. 1979.

[2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[3] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 2, pp. 443–445, 1985.

[4] G. Kim, Y. Lu, Y. Hu, and P. C. Loizou, "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," *J. Acoust. Soc. Am.*, vol. 126, no. 3, pp. 1486–1494, Sept. 2009.

[5] G. Kim and P. C. Loizou, "Improving speech intelligibility in noise using environment-optimized algorithms," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 8, pp. 2080–2090, Nov. 2010.

[6] K. Han and D. L. Wang, "An SVM based classification approach to speech separation," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 4632–4635.

[7] K. Han and D. L. Wang, "Towards generalizing classification based speech separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 1, pp. 168–177, Jan. 2013.

[8] Y. Wang and D. L. Wang, "Boosting classification based speech separation using temporal dynamics.," in *Proc. Interspeech Conf.*, 2012.

[9] Y. Wang and D. L. Wang, "Towards scaling up classification-based speech separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 7, pp. 1381–1390, July 2013.

[10] A. A. Kressner, D. V. Anderson, and C. J. Rozell, "A novel binary mask estimator based on sparse approximation," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2013.

[11] D. L. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, P. Divenyi, Ed., pp. 181–197. Kluwer Academic, 2005.

[12] D. S. Brungart, P. S. Chang, B. D. Simpson, and D. L. Wang, "Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation," *J. Acoust. Soc. Am.*, vol. 120, pp. 4007–4018, 2006.

[13] N. Li and P. C. Loizou, "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction," *J. Acoust. Soc. Am.*, vol. 123, no. 3, pp. 1673–1682, Mar. 2008.

[14] U. Kjems, J. B. Boldt, M. S. Pedersen, T. Lunner, and D. L. Wang, "Role of mask pattern in intelligibility of ideal binary-masked noisy speech," *J. Acoust. Soc. Am.*, vol. 126, no. 3, pp. 1415–1426, Sept. 2009.

[15] D. Byrne, H. Dillon, K. Tran, S. Arlinger, K. Wilbraham, R. Cox, B. Hayerman, R. Hetu, J. Kei, C. Lui, J. Kiessling, M. N. Kotby, N. H. A. Nasser, W. A. H. El Kholy, Y. Nakanishi, H. Oyer, R. Powell, D. Stephens, , T. Sirimanna, G. Tavartkiladze, G. I. Frolenkov, S. Westerman, and C. Ludvigsen, "An international comparison of long-term average speech spectra," *J. Acoust. Soc. Am.*, vol. 96, no. 4, pp. 2108–2120, Oct. 1994.

[16] ITU-T, "Objective measurement of active speech level," Recommendation P.56, International Telecommunications Union (ITU-T), Mar. 1993.

[17] ITU-T, "Artificial voices," Standard P.50, International Telecommunications Union (ITU-T), Sept. 1999.

[18] D. Chan, A. Fourcin, D. Gibbon, B. Granstrom, M. Huckvale, G. Kokkinakis, K. Kvale, L. Lamel, B. Lindberg, A. Moreno, J. Mouropoulos, F. Senia, I. Trancoso, C. Veld, and J. Zeiliger, "EUROM - a spoken language resource for the EU," in *Proc. European Conf. on Speech Communication and Technology*, Sept. 1995, pp. 867–870.

[19] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, Sept. 2011.

[20] S. Gonzalez and M. Brookes, "Speech active level estimation in noisy conditions," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, May 2013.

[21] 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.

[22] M. Brookes, "VOICEBOX: A speech processing toolbox for MATLAB," `http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html`, 1997–2013.

[23] S. Gonzalez and M. Brookes, "Sibilant speech detection in noise," in *Proc. Interspeech Conf.*, Portland, Sept. 2012.

[24] B. C. J. Moore and B. R. Glasberg, "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns," *J. Acoust. Soc. Am.*, vol. 74, pp. 750–753, 1983.

[25] 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.

[26] L. Breiman, J. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*, Chapman and Hall, Jan. 1984.

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

[28] D. L. Wang, U. Kjems, M. S. Pedersen, J. B. Bolt, and T. Lunner, "Speech perception of noise with binary gains," *J. Acoust. Soc. Am.*, vol. 124, no. 4, pp. 2303–2307, Oct. 2008.

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

[30] ITU-T, "Test signals for use in telephonometry," Recommendation P.501, International Telecommunications Union (ITU-T), Aug. 1996.

[31] Y. Hu and P. C. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," *Speech Communication*, vol. 49, no. 7–8, pp. 588–601, July 2007.