EFFICIENT ESTIMATION OF INTER-SUBBAND SPEECH CORRELATIONS

Alexander Schasse, Rainer Martin

Institute of Communication Acoustics
Ruhr-Universität Bochum
44780 Bochum, Germany
e-mail: {firstname.lastname}@rub.de

Ulrich Kornagel, Eghart Fischer, Henning Puder*

Sivantos GmbH
Gebbertstrasse 125
91058 Erlangen, Germany
e-mail: {firstname.lastname}@sivantos.de

ABSTRACT

We propose an approach to compute the inter-subband correlation (ISBC) of noisy speech signals to distinguish between speech and noise segments in the time-frequency plane. The proposed spectral correlation estimator provides information about the input signal which can be used to derive a binary mask or the speech-presence probability. Unlike other approaches it does not require an estimate of the noise power. To this end we analyse a received noisy speech signal in the modulation domain and identify similarly modulated subband signals within a range of modulation frequencies that are typical for speech signals. Based on this pre-processing step, we identify a single reference subband that most likely contains speech and estimate the spectral correlations with respect to this reference band. The algorithm proposed in this paper aims at a very low computational complexity which makes it suitable for hearing aids.

***Index Terms*—** modulation spectrum, spectral correlation, noise reduction, binary mask, speech presence probability

1. INTRODUCTION

Single-channel noise-reduction (NR) algorithms are often implemented in the short-time frequency domain and usually apply instantaneous and real-valued weights to estimate the clean speech DFT coefficients in each frequency subband. These gains generally depend on the signal-to-noise ratio (SNR) observed in each subband as, for instance, in case of the Wiener filter [1, 2], the spectral subtraction algorithm [3], or MMSE amplitude estimators [4, 5, 6]. The transformation to the Fourier domain resolves short-term correlations within the current signal frame. However, since the span of correlation, especially of voiced speech sounds, is larger than the frame length, we still observe significant correlation in each subband and even more between different subbands. This *inter-subband correlation* (ISBC) provides additional information about speech activity especially when it is evaluated in the modulation domain. It is known that synchronous temporal variations of the envelope of audio signals extracted in different frequency bands represent important auditory cues [7, 8, 9] and have, as shown e.g., in [10], great impact on speech intelligibility. When changing the temporal envelope, the intelligibility of processed signals is affected, as analyzed for instance, in [11, 12, 13]. Thus, in a noisy or reverberant speech signal synchronous temporal modulations and features derived from the amplitude modulation spectrum may be used to separate speech from interference as successfully demonstrated, e.g., in the context of automatic speech recognition [14, 15] or speech segregation [16, 17].

In this paper we investigate the ISBC by means of an efficient computational algorithm for the detection of similarly modulated subband signals and thus of speech presence. The correlation between subbands can be used, either to estimate the ideal binary mask (IBM) or to derive a soft mask in the sense of a speech-presence probability (SPP) estimator. Both types of masks can be applied directly to the noisy signal, or they can be used as an additional source of information for existing and novel speech enhancement algorithms. Unlike existing SPP estimation methods, we do not require an estimate of the noise power spectral density or a local SNR.

The remainder of this paper is structured as follows. In Section 2, we introduce the signal model and the IBM which is used as a reference in the following investigations. Section 3 then describes the estimation of ISBCs in the modulation domain. Beside the possibility to derive a correlation-based binary mask, Section 4 describes an ISBC-based SPP estimator. Finally, we provide evaluation results and a discussion of results in Section 5 and Section 6, respectively.

2. SIGNAL MODEL AND IDEAL BINARY MASK

To derive the inter-subband correlation (ISBC) in the context of single-channel speech signal processing, we assume an additive noise model in the short-time Fourier domain (STFT), i.e., we have

$$Y(k,m) = X(k,m) + V(k,m), \tag{1}$$

where $Y(k,m)$ is the STFT of the noisy speech, $X(k,m)$ is the STFT of the clean speech component, and $V(k,m)$ is the STFT of the additive noise. The indices k and m indicate the frequency bin and the time frame, respectively. An SNR-based IBM assumes the clean speech signal power as well as the additive noise power to be known. Then, we can define the IBM via a local criterion as

$$\mathcal{IBM}(k,m) = \begin{cases} 1 & \text{, if SNR}(k,m) \geq \delta_{\text{IBM}} \\ 0 & \text{, otherwise} \end{cases}, \tag{2}$$

where SNR(k,m) represents the true (local) SNR as prior information in each time-frequency (TF) bin and δ_{IBM} is a threshold. The IBM is a widely used concept in single and multi-channel source separation [18] and computational auditory scene analysis (CASA) [19, 20, 21], especially since it is known to improve speech intelligibility when applied directly to mixed signals. Significant effort has been directed towards a batch or online estimation of the IBM to achieve noise reduction or speech segregation, e.g. [22, 23, 24, 17]. Here we use the IBM as an ideal reference only.

*Ulrich Kornagel is now with Technische Hochschule Nuremberg, Germany. E-mail: ulrich.kornagel@th-nuernberg.de

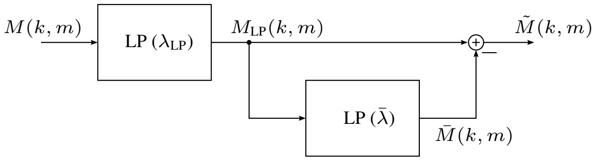


Fig. 1: Extraction of modulation components: In each subband, envelopes $M(k,m)$ are first low-pass filtered which results in $M_{\mathrm{LP}}(k,m)$. Then, a high-pass filter is applied to remove the mean $\tilde{M}(k,m) = M_{\mathrm{LP}}(k,m) - \bar{M}(k,m)$.

3. EFFICIENT COMPUTATION OF INTER-SUBBAND CORRELATIONS

The concept of the ISBC as proposed in this work is basically to detect similarly modulated subband signals, based on the assumption that the strongest modulation components originate from the target signal. The ISBC computation process can be divided into the following steps [25]:

1. After the STFT, the noisy input signal is analyzed in the modulation domain. We focus on modulation frequencies that are most representative of speech.
2. Identification of a reference subband which most likely contains speech.
3. Computation of the cross-correlation of each subband and the selected reference subband in the modulation domain with respect to a predefined or adaptive memory length.

The remainder of this section provides a more detailed description of these processing steps.

3.1. Preprocessing

We aim to identify similarly modulated subband signals in the noisy input signal $Y(k,m)$ with low complexity and low latency. Therefore we extract the envelope via an incoherent approach using the log-magnitude of the input signal,

$$M(k,m) = 10\log_{10}|Y(k,m)|^2\ . \tag{3}$$

We note that more accurate incoherent and coherent envelope detectors are available as proposed, e.g., in [26, 27, 16]. However, they introduce additional delay, require the estimation of the instantaneous frequency and are often sensitive to noise.

To extract envelope modulations that are typical for speech, i.e., within a modulation frequency range of 5 - 25 Hz, the log-envelope $M(k,m)$ is then filtered using a band-pass filter and the resulting signal is denoted as $\tilde{M}(k,m)$. We like to keep the computational complexity as low as possible and therefore apply a band-pass filter consisting of two simple first-order recursive systems with parameters λ_{LP} and $\bar{\lambda}$ as shown in Fig. 1.

3.2. ISBC Computation

Since the computation of the correlation between all subband combinations results in a high computational complexity, we propose to define a single, but data-dependent reference subband, identified by its frequency bin index $k_{\mathrm{ref}}(m)$. This reference subband index should be chosen such that it contains the target signal with high probability. In our experiments, the most effective selection scheme relies on the filtered log-envelope

$$k_{\mathrm{ref}}(m) = \arg_k \max \tilde{M}(k,m). \tag{4}$$

With respect to this reference subband, the ISBC is defined as

$$\mathrm{ISBC}(k,m) = \frac{\mathrm{E}\left\{\tilde{M}(k_{\mathrm{ref}}(m),m)\tilde{M}(k,m)\right\}}{\sqrt{\mathrm{E}\left\{\left|\tilde{M}(k_{\mathrm{ref}}(m),m)\right|^2\right\}\mathrm{E}\left\{\left|\tilde{M}(k,m)\right|^2\right\}}}, \tag{5}$$

where in practice all statistical expectations $\mathrm{E}\{\bullet\}$ are approximated with recursive temporal averages of first order and smoothing parameter λ_{corr}. This parameter determines the correlation memory. The choice of the reference subband defined in (4) results in a high contrast between speech activity and speech pauses. During speech activity it indicates the most dominant speech components, while during speech pauses it shows a random characteristic. Especially this latter effect is important for the overall performance of the method as it triggers a quick decay of the correlation estimate at speech offsets. Figure 2 shows a short segment of speech in traffic noise at 0 dB seg. input SNR and the corresponding ISBC (Fig. 2 (f)). By thresholding the ISBC a binary mask may be created. In Fig 2 (e) and (g) we depict two versions of a binary mask: $\mathrm{ISBC}_{\mathrm{BM(0.1)}}$ is a mask which sets all TF-bins to one for which we have ISBC $\geq \delta_{\mathrm{ISBC}}$ with $\delta_{\mathrm{ISBC}} = 0.1$ and to zero otherwise. Correspondingly, $\mathrm{ISBC}_{\mathrm{BM(0.25)}}$ in Fig 2 (g) uses a slightly larger correlation level of $\delta_{\mathrm{ISBC}} = 0.25$.

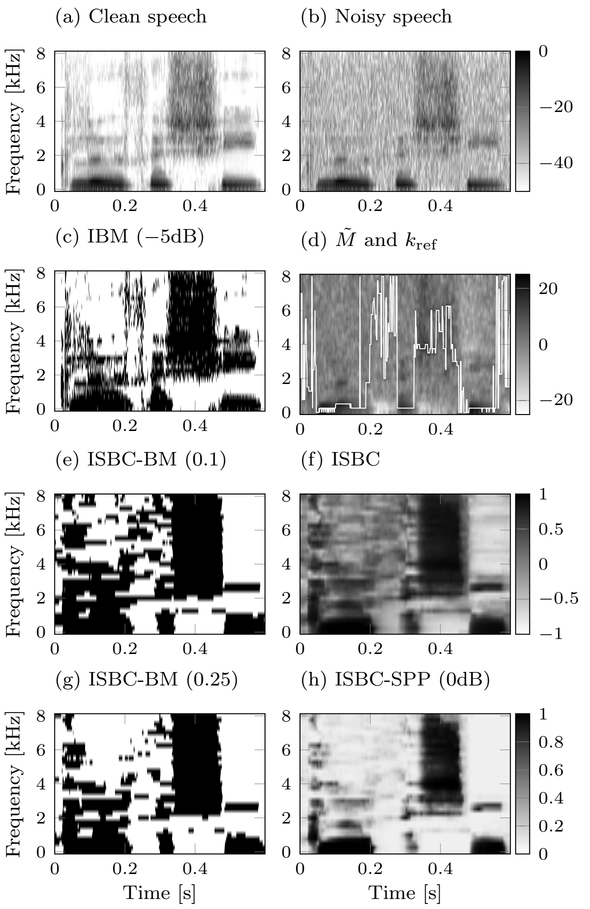
4. ISBC-BASED SPEECH-PRESENCE PROBABILITY

While the ISBC can be used to derive a simple binary mask based on a hard decision with a (possibly frequency-dependent) threshold $\delta_{\mathrm{ISBC}}(k)$, it is more promising to derive an estimate of the speech-presence probability given the noisy input signal and its ISBC. For this purpose, we define the two hypotheses

$$\Theta = \{\theta_0 : \text{“Speech and Noise”}, \theta_1 : \text{“Noise Only”}\} \tag{6}$$

in terms of a binary random variable. The *a-priori* probabilities $P(\Theta = \theta_0)$ and $P(\Theta = \theta_1)$ are estimated as the relative frequencies of ones and zeros in the IBM (2). Based on this binary random variable, we interpret the ISBC as a random variable z on the interval $[-1, 1]$, defined by conditional probability density functions (PDF) $p_{\mathrm{ISBC}|\Theta}(z|\Theta = \theta_0)$ and $p_{\mathrm{ISBC}|\Theta}(z|\Theta = \theta_1)$. We estimate these PDFs using histograms of the ISBC for many noisy speech signals (8 different speech signals taken from the TSP database [28], 12 different noise signals taken from the SoundIdeas6000 database [29], 3 different input SNRs). We use the IBM (2) to distinguish between “Speech and Noise” and “Noise Only”, and collect ISBC values for both classes. The histograms shown in Fig. 3 are estimates of these conditional PDFs.

We find that large correlation values are more likely to represent speech. Noise is represented by a more or less symmetrical histogram with its maximum between $z = -0.2$ and $z = 0$. For $\Theta = \theta_0$ and ISBC values near $z = 1$ the histogram shows a decline which is caused by insufficient coverage of very high correlation values in the training data. This effect may be ignored in the PDF model. Thus, based on the histograms shown in Fig. 3, we model the


(a) Clean speech

(b) Noisy speech

(c) IBM (−5dB)

(d) $\tilde{M}$ and k_{ref}

(e) ISBC-BM (0.1)

(f) ISBC

(g) ISBC-BM (0.25)

(h) ISBC-SPP (0dB)

Fig. 2: Spectrograms for a short speech segment in traffic noise mixed at 0 dB seg. input SNR. (a) clean speech signal, (b) noisy speech signal, (c) IBM with $20\lg(\delta_{\text{IBM}}) = -5$ dB, (d) modulation components and selected reference band (indicated as a white line), (e) binary mask via thresholding the ISBC at $\delta_{\text{ISBC}} = 0.1$, (f) ISBC, (g) binary mask via thresholding the ISBC at $\delta_{\text{ISBC}} = 0.25$, (h) SPP $P_{\Theta|\text{ISBC}}(\Theta = \theta_0|z)$ for an SPP ratio $R_{\text{SPP}} = 6.1$.

ISBC via a beta-distribution in the interval $[a, b]$,

$$p(z|\alpha, \beta, a, b) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}(b-a)^{-(\alpha+\beta-1)}(z-a)^{\alpha-1}(b-z)^{\beta-1}, \qquad (7)$$

where $\Gamma(t) = \int_0^\infty x^{t-1}\mathrm{e}^{-x}dx$ is the complete Gamma function. Table 1 summarizes the parameter values we use in our experiments.

4.1. Estimation of Speech Presence Probability

Using the *a-priori* SPP $P(\Theta = \theta_0) = 1 - P(\Theta = \theta_1)$ as well as the conditional distributions $p_{\text{ISBC}|\Theta}(z|\Theta = \theta_0)$ and $p_{\text{ISBC}|\Theta}(z|\Theta = \theta_1)$, we can now define the posterior probability of "Speech and Noise" following Bayes rule

$$P_{\Theta|\text{ISBC}}(\Theta = \theta_0|z) = \frac{p_{\text{ISBC}|\Theta}(z|\Theta = \theta_0)P(\Theta = \theta_0)}{p_{\text{ISBC}|\Theta}(z|\Theta = \theta_0)P(\Theta = \theta_0) + p_{\text{ISBC}|\Theta}(z|\Theta = \theta_1)P(\Theta = \theta_1)}, \qquad (8)$$

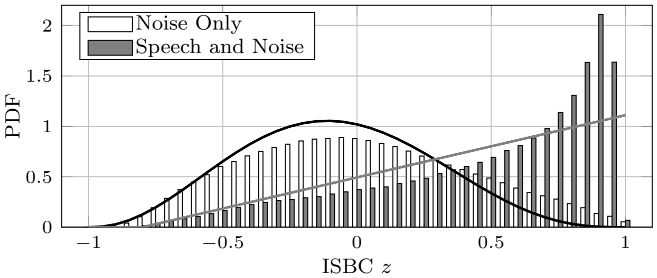


Fig. 3: ISBC histograms (bars) for the classes "Speech and Noise" and "Noise Only" and the respective fits using the Beta distribution (solid lines) for 0 dB seg. input SNR.

which defines the speech-presence probability for a given ISBC value z. This distribution basically describes a mapping of ISBC values to a range of values between 0 and 1 with respect to the conditional distributions and the *a-priori* SPP. Figure 4 shows this mapping when using the fitted beta distributions, as well as the histogram data for a fixed *a-priori* SPP of $P(\Theta = \theta_0) = 0.5$. We note that the mapping looks very similar for -10, 0, and 10 dB seg. input SNR. For the particular parametrization given in Table 1, the ISBC-based SPP may therefore be approximated by

$$P_{\Theta|\text{ISBC}}(\Theta = \theta_0|z) = \frac{1}{1 + C\tfrac{(z+1)^{2.4}(1-z)^3}{z+0.8}R_{\text{SPP}}}, \qquad (9)$$

labeled in Fig. 4 as 'Beta Distribution Fit'. Here, C summarizes constant terms of the beta distributions and evaluates to

$$C = \frac{\Gamma(7.4)}{\Gamma(3.4)\Gamma(4)}\frac{\Gamma(2)\Gamma(1)}{\Gamma(3)}\frac{1.8^2}{2^{6.4}} \approx 1.6530\,. \qquad (10)$$

$R_{\text{SPP}} = P(\Theta = \theta_1)/P(\Theta = \theta_0)$ is the *a-priori* SPP ratio which may be used to bias the SPP towards one of the two hypotheses.

5. EXPERIMENTAL RESULTS

To evaluate the proposed algorithm we process 180 seconds of speech (6 female and 6 male speakers [28]) and 6 types of babble and traffic noise [29]. All signals are sampled at a rate of 16 kHz and transformed to the STFT domain via a 64-point FFT and a frame advance of 16 samples. We apply two version of binary masks (ISBC thresholds of 0.1 and 0.25) and the soft speech presence probability, all of which are derived from the ISBC data, to the time-frequency representation of the noisy signals. After applying these masks we use an IDFT and the overlap-add method to reconstruct an enhanced time-domain signal. We evaluate the quality of the reconstructed signal in terms of PESQ in comparison to the Ideal Binary Mask (IBM) and furthermore compute the mean-square error (MSE) between the IBM and the hard and soft masks derived from the ISBC

parameter	"Speech and Noise"	"Noise Only"
minimum value a	-0.8	-1
maximum value b	1	1
shape parameter α	2	3.4
shape parameter β	1	4

Table 1: Parameters of the Beta distribution used to model the ISBC for classes "Speech and Noise" and "Noise Only".

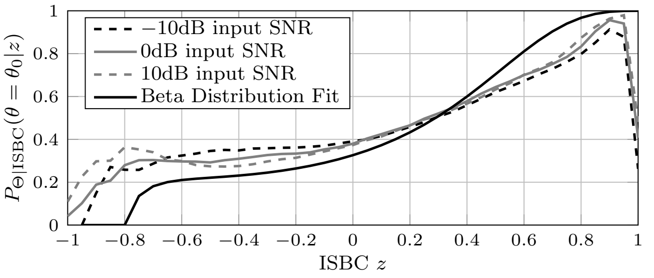


Fig. 4: Posterior distribution which defines the SPP for a given ISBC value for a fixed *a-priori* SPP of $P(\Theta = \theta_0) = 0.5$. The gray lines as well as the black dashed line indicate the posterior distributions using the histogram data collected on a large database. The solid black line represents the manual fit of the beta distribution.

Table 2: PESQ MOS-LQO scores averaged over 12 speaker and for different noise types [29] at 0 dB SNR.

noise type	noisy input	IBM	ISBC-SPP
white Gaussian	1.0793	1.8116	1.2968
light traffic, dry road	1.2007	1.5895	1.3483
medium traffic, dry road	1.2118	1.7912	1.4544
heavy traffic, dry road	1.1832	1.7479	1.3892
city traffic, horns, rumble	1.1615	1.7104	1.3892
bar/pub, medium crowd	1.2467	1.8788	1.2738
open kitchen, large crowd	1.1803	1.8366	1.2234
restaurant, small crowd	1.2007	1.7640	1.3210

Table 3: PESQ MOS-LQO scores averaged over 12 speaker and for different noise types [29] at 10 dB SNR

noise type	noisy input	IBM	ISBC-SPP
white Gaussian	1.4837	2.2942	1.9480
light city traffic	1.7818	2.1730	2.0988
medium city traffic	1.8905	2.4314	2.2986
heavy city traffic	1.7010	2.2559	2.1414
city traffic, rumble	1.7992	2.2789	2.2173
bar/pub, medium crowd	1.8720	2.4459	2.0693
open kitchen, large crowd	1.7391	2.3792	1.9413
restaurant, small crowd	1.8555	2.3547	2.1599

in the short-time frequency domain. The MSE is defined as

$$\mathrm{MSE} = \frac{1}{KM} \sum_{k} \sum_{m} (\mathcal{IBM}(k,m) - \mathrm{ISBC}_{*}(k,m))^2 \qquad (11)$$

where ISBC_{*} is one of $\mathrm{ISBC}_{\mathrm{BM}(0.1)}$, $\mathrm{ISBC}_{\mathrm{BM}(0.25)}$, or $\mathrm{ISBC}_{\mathrm{SPP}}$.

Table 2 depicts the PESQ scores for the IBM and the proposed ISBC-SPP method for several noise types at an input SNR of 0 dB. Although the proposed method is able to improve the PESQ scores and thus the predicted quality of the signals it does not come close to the performance of the IBM. For the SNR of 10 dB shown in Table 2 the ISBC method shows a substantial improvement. For noise types with a broadband random spectrum, e.g. 'city medium' and 'traffic light', the proposed ISBC-SPP approaches the performance of the IBM. For noise types with clearly audible background voices less improvements are achieved.

In Figure 5 we depict the mean square error of the estimated masks with respect to the IBM. We find that the application of the speech presence probability leads to a noticable reduction of the error. This effect is also observed in the auditory quality of the processed signals: In an informal listening test, the soft-mask results in a much better quality than the binary mask as spectral outliers and fluctuations are less noticable.

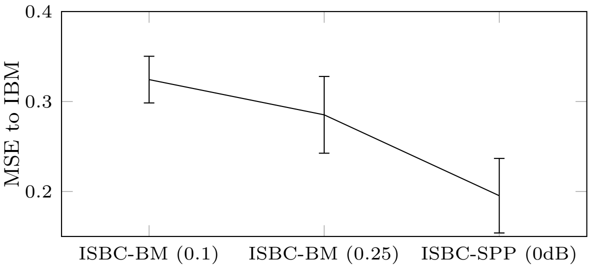


Fig. 5: Mean-squared error between IBM and masks derived from ISBC for 60s of noisy speech at 0 dB SNR. ISBC-BM (0.1): A hard threshold at correlation level 0.1 was applied to the ISBC. ISBC-BM (0.25): A hard threshold at correlation level 0.25 was applied to the ISBC. ISBC-SPP (0 dB): SPP computed at a global SNR of 0 dB.

6. DISCUSSION AND OUTLOOK

This paper introduces an efficient algorithm to analyze spectral speech correlations in the modulation domain in order to gain information about the noisy speech signal for single-channel noise reduction and speech detection purposes. Based on this correlation analysis, we derive a speech-presence probability estimator which, unlike established methods, does not require an estimate of the noise power in each subband. The inter-subband correlation and the resulting binary and soft ISBC-SPP masks provide a noticable separation between speech and noise in the short-time Fourier domain.

When we apply the estimated soft mask directly to noisy speech signals we find a significant attenuation of the noise and hence also an improvement of PESQ MOS-LQO scores as reported in Tables 2 and 3. However, especially in low SNR conditions there are also processing artifacts which originate from estimation errors and from a delayed reaction to speech onsets and speech offsets. These artifacts lead to a small reduction in the predicted intelligibility as for instance computed with the STOI (Short-Time Objective Intelligibility [30]) measure (not reported here). Thus, the brute-force application of the derived masks do currently not provide the auditory quality of a well-tuned conventional noise reduction filter, e.g. based on an online noise power estimate and the Wiener filter, as discussed, e.g. in [31, 1, 2, 32]. Nevertheless, it provides interesting insights into the spectro-temporal composition of the signal. As it is entirely different from conventional methods used in hearing aids and has a lower computational complexity compared to methods using more elaborate statistical or auditory models, we believe that it constitutes a useful addition to the tools of the trade.

In future works we will consider improved methods for the selection of the reference band as the temporal dynamics of the selection process has a significant influence on the variations of the estimated ISBC. Also, a major share of the processing latency of the approach resides in the computational memory of the correlator. The tradeoff between an efficient reference band selection, an efficient computation of the correlation coefficients, and the variance of the estimated masks will be subject to further investigations.

7. REFERENCES

[1] P. Vary and R. Martin, *Digital Speech Transmission: Enhancement, Coding And Error Concealment*, John Wiley & Sons, 2006.

[2] C. Breithaupt and R. Martin, "Noise reduction – Statistical analysis and control of musical noise," in *Advances in Digital Speech Transmission*, Rainer Martin, Ulrich Heute, and Christiane Antweiler, Eds., pp. 107–133. John Wiley & Sons, 2008.

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

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

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

[6] C. Breithaupt, M. Krawczyk, and R. Martin, "Parameterized MMSE Spectral Magnitude Estimation for the Enhancement of Noisy Speech," in *IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, 2008, pp. 4037–4040.

[7] A.S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*, MIT Press, 1990.

[8] G.J. Brown and M. Cooke, "Temporal Synchronization in a Neural Oscillator Model of Primitive Auditory Stream Segregation," in *Computational Auditory Scene Analysis*, D.F. Rosenthal and H.G. Okuno, Eds., pp. 87 – 103. 1998.

[9] A.M. Liberman, *Speech: A Special Code*, A Bradford book. MIT Press, 1996.

[10] R. Drullman, "Temporal envelope and fine structure cues for speech intelligibility," *J. Acoust. Soc. Am.*, vol. 97, no. 1, pp. 585–592, Jan. 1995.

[11] R. Drullman, J.M. Festen, and R. Plomp, "Effect of temporal envelope smearing on speech reception," *J. Acoust. Soc. Am.*, vol. 95, pp. 1053–1064, Feb. 1994.

[12] R. Drullman, J. M. Festen, and R. Plomp, "Effect of reducing slow temporal modulations on speech reception," *J. Acoust. Soc. Am.*, vol. 95, no. 5, pp. 2670–2680, May 1994.

[13] R.A. van Buuren, J.M. Festen, and T. Houtgast, "Compression and expansion of the temporal envelope: Evaluation of speech intelligibility and sound quality," *J. Acoust. Soc. Am.*, vol. 105, no. 5, pp. 2903–2913, May 1999.

[14] H. Hermansky, N. Morgan, A. Bayya, and P. Kohn, "RASTA-PLP Speech Analysis Technique," in *IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, 1992, vol. I, pp. 121–124.

[15] H. Hermansky and N. Morgan, "RASTA Processing of Speech," *IEEE Trans. Speech and Audio Process.*, vol. 2, no. 4, pp. 578–589, Oct 1994.

[16] S.M. Schimmel, L.E. Atlas, and K. Nie, "Feasibility of Single Channel Speaker Separation Based on Modulation Frequency Analysis,," in *IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, 2005, p. 221224.

[17] T. May and T. Dau, "Computational speech segregation based on an auditory-inspired modulation analysis," *J. Acoust. Soc. Am.*, vol. 136, no. 6, pp. 3350 – 3359, 2014.

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

[19] G. Hu and D. Wang, "Speech segregation based on pitch tracking and amplitude modulation," in *IEEE Workshop Applications of Signal Process. to Audio and Acoustics (WASPAA)*, New Paltz, NY, U.S.A., Oct. 2001, pp. 79–82.

[20] N. Roman, D. Wang, and G. J. Brown, "Speech segregation based on sound localization," *J. Acoust. Soc. Am.*, vol. 114, no. 4, pp. 2236–2252, Oct. 2003.

[21] D. Wang and G.J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, Wiley-IEEE Press, 2006.

[22] Y. Li and D. Wang, "On the optimality of ideal binary time-frequency masks," *Speech Communication*, vol. 51, pp. 230–239, 2009.

[23] N. Roman and J. Woodruff, "Intelligibility of reverberant noisy speech with ideal binary masking," *J. Acoust. Soc. Am.*, vol. 130, no. 4, pp. 2153 – 61, 2011.

[24] E.W. Healy, S.E. Yoho, Y. Wang, and D. Wang, "An algorithm to improve speech recognition in noise for hearing-impaired listeners," *J. Acoust. Soc. Am.*, vol. 134, no. 4, pp. 3029 – 38, 2013.

[25] A. Schasse, *Single-Channel Noise Reduction based on Long-Term Speech Correlations with Application to Hearing Aids*, Ph.D. thesis, Institute of Communication Acoustics, Ruhr-Universität Bochum, 2016.

[26] James F. Kaiser, "On a simple algorithm to calculate the 'energy' of a signal," in *IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, Albuquerque, New Mexico, U.S.A., Apr. 1990, pp. 381–384 vol.1.

[27] L. Atlas and C. Janssen, "Coherent modulation spectral filtering for single-channel music source separation," in *IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, Philadelphia, Pennsylvania, U.S.A., Mar. 2005, pp. iv/461 – iv/464 Vol. 4.

[28] P. Kabal, "TSP speech database, version 1.0," Tech. Rep., Telecommunications & Signal Processing Laboratory, McGill University, Montreal, Canada, 2002.

[29] Sound Ideas, "Sound ideas 6000 database," 2002, Electrical & Computer Engineering, McGill University.

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

[31] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Audio, Speech and Language Process.*, vol. 9, no. 5, pp. 504–512, Jun. 2001.

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