# PITCH-BASED NON-INTRUSIVE OBJECTIVE INTELLIGIBILITY PREDICTION

*Charlotte Sørensen*[1,2], *Angeliki Xenaki*[2], *Jesper B. Boldt*[2] *and Mads G. Christensen*[1]

[1]Audio Analysis Lab, AD:MT, Aalborg University, Denmark
[2]GN Hearing A/S, Lautrupbjerg 7, DK-2750, Ballerup, Denmark
{csoerensen,axenaki,jboldt}@gnresound.com, {mgc}@create.aau.dk

## ABSTRACT

Automatic adjustment of the hearing aid according to the intelligibility for the user in the environment could be beneficial. While most intelligibility metrics require a clean speech reference, i.e. intrusive methods, this is rarely available in real-life. This paper proposes a non-intrusive intelligibility metric in which a reconstruction of the clean speech is used in the established intrusive short-time objective intelligibility (STOI) metric. The reconstruction of the clean speech is based on pitch-features of the desired source using a spatio-temporal harmonic model. This model takes advantage of both the spatial and spectral separation of the desired source and interferers to reconstruct the clean signal. The simulations show a high correlation between the proposed pitch-based STOI (PB-STOI) and the original intrusive STOI and hence is promising for online processing of intelligibility.

***Index Terms***— Pitch estimation, non-intrusive objective intelligibility prediction, hearing aids

## 1. INTRODUCTION

One of the main issues encountered by hearing aid (HA) users is severely degraded speech intelligibility in noisy multi-talker environments such as the ”cocktail party problem” [1, 2]. Generally, the speech intelligibility for users of assistive listening devices depends highly on the specific listening environment. As such, additional speech enhancement processing may be beneficial in some listening environments whereas the exact same algorithms can have a negative impact on the quality and intelligibility in other listening environments [3, 4]. In HA technology, automatic intelligibility assessment of the listening environment would be beneficial for the user such that speech enhancement is only applied when necessary [5, 6]. This could be facilitated by an online intelligibility evaluation of the listening environment. Thus, it could be beneficial if objective intelligibility metrics could be used in the online processing of HAs.

There are various intrusive methods to predict the speech intelligibility with acceptable reliability such as the short-time objective intelligibility (STOI) metric [7] and the normalized covariance metric (NCM) [8]. However, these methods are intrusive, i.e., they all require access to the clean-speech reference which is rarely available in real-life. A number of non-intrusive methods have been introduced that do not require access to the clean speech signal, e.g. the modulation spectrum area (ModA) [9] or the speech-to-reverberation modulation energy ratio (SRMR) [10]. However, both of these non-intrusive measures are limited to the assessment of reverberated speech signals and are still inferior to the intrusive measures according to a recent review [6].

This paper proposes a method that non-intrusively estimates the speech intelligibility in the listening environment for HAs. Similar to the approaches in [11, 12] a prediction of the speech intelligibility is obtained by comparing a reconstruction of the clean speech with the noisy speech using an established and reliable intrusive framework, e.g. STOI [6, 13]. The clean speech is obtained by estimating relevant signal features assuming the desired source consists of a number of narrowband signals with harmonically related carrier frequencies using a spatio-temporal model. Combining spatial (i.e. direction of arrival) and temporal (i.e. pitch) cues improves the accuracy of the reconstruction as it resolves ambiguities, e.g. due to reverberation or competing speakers. The proposed method can then potentially be used as an alternative to environment classification by determining, whether the intelligibility is below a certain threshold [14].

## 2. METHOD

In this section the approach behind the PB-STOI metric is presented. A block diagram incorporating the framework is shown in Fig. 1. In the first step, the sound field is recorded with a microphone array. Then, the pitch of the desired speech signal is estimated and the speech is reconstructed using the pitch and direction of arrival of the desired speech signal. Finally, a non-intrusive prediction, d($n$), is given on a 0-1 scale by comparing the correlation of the reconstructed clean speech with the noisy version using the intrusive STOI framework.

This work was supported by the Innovation Fund Denmark, Grant No. 99-2014-1.

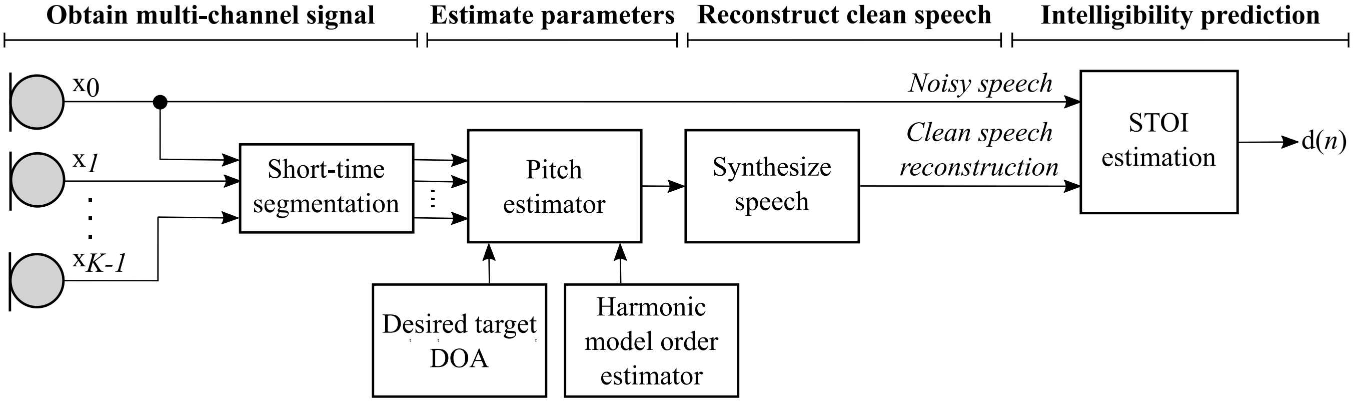


**Fig. 1**. Block diagram of the proposed pitch-based non-intrusive objective intelligibility measure in which reconstruction of the clean speech is obtained using the estimated pitch and compared with the output of an omnidirectional microphone using the original intrusive STOI.

### 2.1. Signal model

A multi-channel spatio-temporal harmonic model is applied based on the model from [15] in order to reconstruct the clean speech signal as input to the intrusive intelligibility metric. In the proposed method it is assumed that $K$ microphones are used to obtain the desired signal added to a mixture of interfering sources and background noise for a frame length of $N$ such for the $k$'th microphone, the data vector $\mathbf{x}_k = [x_k(0)\ x_k(1)\ \ldots\ x_k(N-1)]^T$ for $k = 0,\ldots,K-1$. The desired source is assumed to be periodic, which is an appropriate assumption for short segments of voiced speech [16]. As such, the data vector $\mathbf{x}_k$ can be modeled as

$$\mathbf{x}_k = \beta_k \mathbf{Z}\mathbf{D}(k)\boldsymbol{\alpha} + \mathbf{e}_k, \tag{1}$$

with $\mathbf{Z} = [\mathbf{z}(\omega_0)\ldots\mathbf{z}(L\omega_0)], \mathbf{z}(l\omega_0) = [1\ldots e^{jl\omega_0(N-1)}]$ for $n = 0,\ldots,N-1$, $\mathbf{D}(k) = \mathrm{diag}([e^{-j\omega_0 f_s\tau_k}\ldots e^{-jL\omega_0 f_s\tau_k}])$ for $l = 1,\ldots,L$ with all other entries equal to zero and $\mathbf{e}_k$ is the sum of the recorded noise and interference. Furthermore, $\omega_0$ is the fundamental frequency, $f_{\text{s}}$ is the sampling frequency and $\tau_k$ is the delay of the desired target source between microphone 0 and the $k$'th microphone giving the direction of arrival (DOA). Moreover, $\beta_k$ is the attenuation of the desired source at the $k$'th microphone, $\boldsymbol{\alpha} = [\alpha_1\ldots\alpha_L]^T$ is the complex amplitudes given by $\alpha_l = A_l e^{j\phi_l}$, $L$ is the number of harmonics, $A_l > 0$ and $\phi_l$ are the real amplitude and phase of the $l$'th harmonic, respectively.

### 2.2. Pitch-based intelligibility prediction

The pitch of the desired target source is found by exploiting the spatio-temporal harmonic model structure of the multi-channel signal using the joint pitch and DOA estimation method presented in [15]. In the following, the basic principles and deviations from the original method are explained.

Assuming the noise is uncorrelated white Gaussian with variance $\sigma_k^2$ in each channel, the log-likelihood function of the complex data vector $\mathbf{x}_k$ can be written as [15]

$$\ln p(\mathbf{x}_k;\psi) = \\ -NK\ln\pi - N\sum_{k=0}^{K-1}\ln\sigma_k^2 - \sum_{k=0}^{K-1}\frac{\|\mathbf{e}_k\|^2}{\sigma_k^2} \tag{2}$$

with the vector $\psi$ containing the signal parameters for $\mathbf{x}_k$. Even though this assumption may seem unreasonable the white Gaussian noise distribution maximizes the entropy of the noise and is a good choice for the noise probability density function [15]. Then, the pitch can be estimated by maximizing the log-likelihood function by differentiating with respect to the amplitudes, $\hat{\boldsymbol{\alpha}}$, the attenuation factor, $\beta_k$, and the noise variance, $\sigma_k^2$, respectively. As mentioned in [15] these parameters are dependent on each other and are therefore estimated by initially setting the $\beta_k$'s and $\sigma_k^2$'s to 1 and iterating over the expressions in Equation (3), (4) and (5). The estimated complex amplitudes are given by

$$\hat{\boldsymbol{\alpha}} = \left[\sum_{k=0}^{K-1}\frac{\beta_k^2}{\sigma_k^2}\mathbf{D}^H(k)\mathbf{Z}^H\mathbf{Z}\mathbf{D}(k)\right]^{-1}\sum_{k=0}^{K-1}\frac{\beta_k}{\sigma_k^2}\mathbf{D}^H(k)\mathbf{Z}^H\mathbf{x}_k \tag{3}$$

The estimated attenuation of the desired source at the $k$'th microphone can be obtained as

$$\hat{\beta}_k = \frac{\mathrm{Re}\{\boldsymbol{\alpha}^H\mathbf{D}^H(k)\mathbf{Z}^H\mathbf{x}_k\}}{\boldsymbol{\alpha}^H\mathbf{D}^H(k)\mathbf{Z}^H\mathbf{Z}\mathbf{D}(k)\boldsymbol{\alpha}} \tag{4}$$

Moreover, the noise variance can be found as

$$\hat{\sigma}_k^2 = N^{-1}\|\hat{\mathbf{e}}_k\|^2, \tag{5}$$

where $\hat{\mathbf{e}}_k = \mathbf{x}_k - \beta_k\mathbf{Z}\mathbf{D}(k)\boldsymbol{\alpha}$. The maximum likelihood estimator of the pitch can then be written as

$$\hat{\omega}_0 = \arg\min_{\omega_0\in\Omega_0}\sum_{k=0}^{K-1}\ln\|\mathbf{x}_k - \hat{\beta}_k\mathbf{Z}\mathbf{D}(k)\hat{\boldsymbol{\alpha}}\|^2 \tag{6}$$

where $\Omega_0$ is a set of possible pitch candidates. Contrary to the original method in [15], the DOA of the desired target source is assumed known such that the problem reduces to spatial filtering rather than DOA estimation and the estimation is only performed over a one-dimensional search. This assumption limits computational complexity as well as makes the model more robust against stronger interfering harmonic sources from other directions. Finally, a reconstruction of the clean speech for the $k$'th microphone can be obtained given the estimated pitch, $\omega_0$ and the delay, $\tau$,

$$\hat{\mathbf{s}}_k = \Pi_{\mathbf{ZD}(k)}\mathbf{x}_k \tag{7}$$

with the projection matrix $\Pi_{\mathbf{A}} = \mathbf{A}(\mathbf{A}^H\mathbf{A})^{-1}\mathbf{A}^H$. The reconstructed clean speech signal to be used as input to the non-intrusive objective intelligibility metric is then obtained by summing the estimated signal over all microphone channels

$$\hat{\mathbf{s}} = \frac{1}{K}\sum_{k=0}^{K-1}\hat{\mathbf{s}}_k \tag{8}$$

Alternatively, the variance estimates in (5) can be used to form a weighted estimate.

### 2.3. Experimental methodology

The proposed metric PB-STOI is evaluated using two different microphone array setups: A broadside uniform linear array (ULA) consisting of $K = 10$ microphones and a behind the ear (BTE) HA setup consisting of two bilateral wireless linked HAs with $K = 4$ microphones. The ULA has a microphone spacing of $d = c/f_\text{s}$ and the delay of the desired source between microphone 0 and the $k$'th microphone is given by $\tau_k = kdc^{-1}\sin\theta$, where the wave propagation speed was $c = 343$ m/s. The DOA of the desired source was $\theta = 0°$ and the sampling frequency was $f_\text{s} = 8$ kHz. For the BTE HA setup the spacing between the microphone on each HA was 1 cm and the spacing between the two HAs was 25 cm.

In the experimental evaluation the set of fundamental frequencies was set to the range $\Omega_0 = 100 - 400$ Hz, the model order was estimated using the maximum a posteriori (MAP) criterion [18], the short-time segmentation window block size was 30 ms and reconstructed by overlap-and-add using a Hanning window with 50% overlap. The simulations were performed using a complex multi-talker scenario with 8 interfering speakers (Fig. 2), reverberation (RT60 = 0.3 s) and ambient white noise in a room with dimensions of 10x6x4 m simulated for 2.5 s using the toolbox McRoomSim [17]. The simulations were carried out in three scenarios at SNRs ranging from -20 to 20 dB; a white noise only scenario, one with interferers and white noise and one with interferers, white noise and reverberation. The desired speech was the utterance ”Why were you away a year, Roy” from the voiced corpus in [19] and the interferers were speech samples from the EUROM_1 database of the English sentence corpus [20].

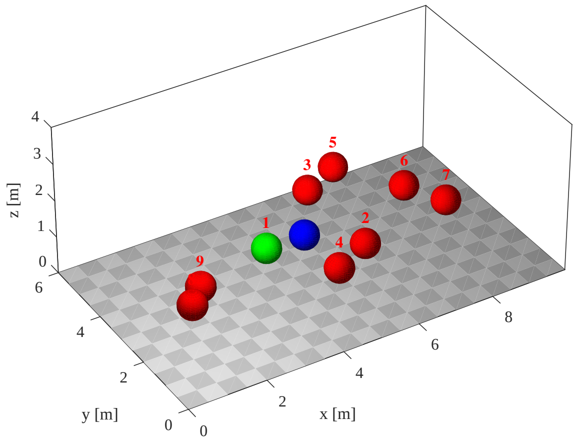


**Fig. 2**. The experimental setup simulated with the software toolbox McRoomSim [17]. The blue, green and red balls illustrate the location of the listener, the desired target source and the interferers, respectively.

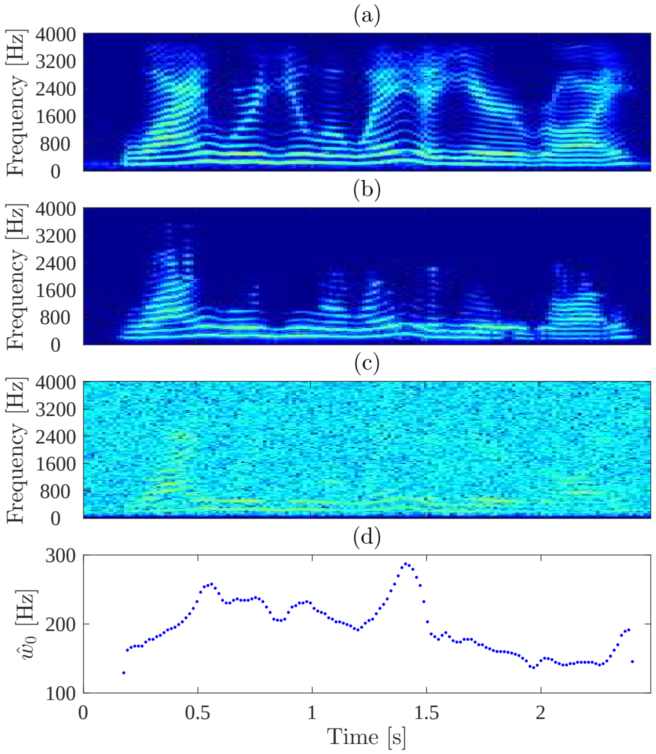
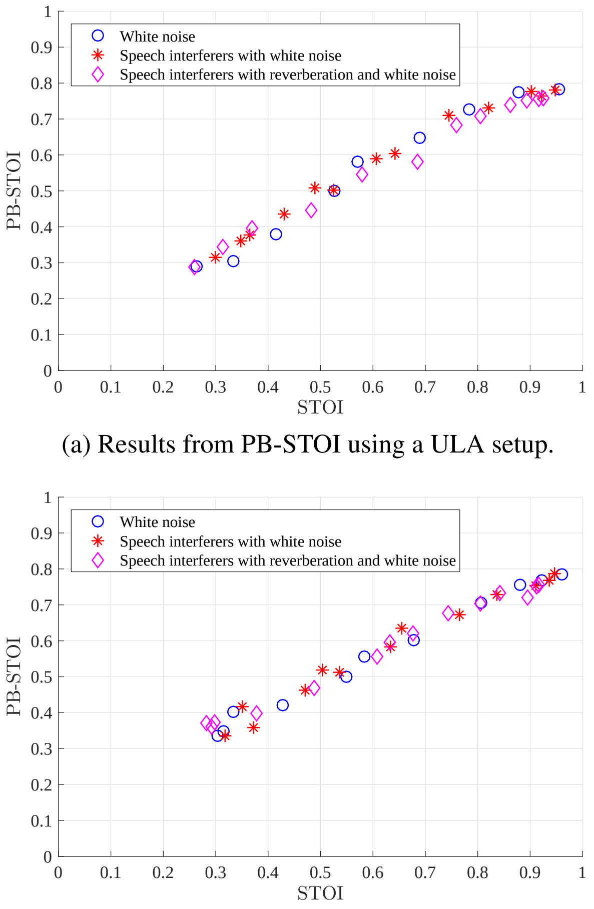


**Fig. 3**. Spectrograms of (a) the clean voiced utterance ”Why were you away a year, Roy”, (b) the reconstructed speech signal using the estimated pitch from the harmonic model, and (c) the noisy signal at 0 dB SNR, and plot of (d) the estimated fundamental frequency from the noisy signal.



(a) Results from PB-STOI using a ULA setup.

(b) Results from PB-STOI using a BTE HA setup.

**Fig. 4**. Scatter plots of the non-intrusive PB-STOI metric versus the intrusive STOI metric. The pitch of the PB-STOI metric is estimated using a multi-channel signal from (a) a ULA with $K = 10$ microphones and (b) two bilateral BTE HAs setup. The circles, asterisks and diamonds show the simulated results for white noise only, multiple interferers with white noise without and with reverberation, respectively.

## 3. RESULTS AND DISCUSSION

The spectrograms of (a) the original clean speech, (b) the equivalent reconstructed signal and (c) the degraded noisy signal at 0 dB as well as (d) the estimated pitch from the noisy signal are depicted in Fig. 3. Comparison of Figs. 3(a) and (b) indicates that the reconstructed speech signal has captured relatively well the features of the original clean signal.

The performance of the proposed intelligibility measure is evaluated by comparing the correlation between the non-intrusive PB-STOI scores against the original intrusive STOI scores in Fig. 4 for (a) the ULA setup and (b) the bilateral BTE HA setup. It can be observed that the PB-STOI scores correlate well with the original intrusive scores with a strong linear trend between the two metrics for both microphone array setups. Thus, it is promising that a small microphone array such as the HA setup can give acceptable results.

**Table 1**. Performance of the proposed metric in terms of Pearson's correlation ($\rho$), the Spearman rank ($\rho_{\text{spear}}$) and Kendall's tau ($\tau$) between PB-STOI and STOI as well as their linear regression lines for a ULA and bilateral BTE HA setup.

| **Setup** | $\rho$ | $\rho_{\text{spear}}$ | $\tau$ | Regression line |
|---|---|---|---|---|
| ULA | 0.9886 | 0.9887 | 0.9287 | $0.74x + 0.11$ |
| BTE HA | 0.9812 | 0.9004 | 0.9922 | $0.67x + 0.16$ |

The performance of the proposed PB-STOI metric is evaluated in Table 1 using three performance criteria often used for assessing objective intelligibility metrics [6, 11]. Pearson's correlation ($\rho$) quantifies the linear relationship, while Spearman's rank ($\rho_{\text{spear}}$) and Kendall's tau ($\tau$) characterize the ranking capability. The values are close to one for all performance criteria indicating high correlation between the intrusive and non-intrusive metric. Hence, the proposed non-intrusive PB-STOI metric can offer a comparable performance to the original intrusive intelligibility metric.

Compared with the study in [11] which uses a similar approach for non-intrusive intelligibility prediction, the proposed PB-STOI metric only requires a calibration of the conversion between PB-STOI and STOI scores depending on the array configuration without any training to the data. However, the experimental evaluation only contained voiced speech and should also be tested on utterances containing unvoiced parts. This could be done by only assessing the intelligibility in the voiced parts of the speech using a voiced speech detector. It is expected to obtain similar results for sentences also containing unvoiced parts, since the most energetic regions occur during the voiced parts. According to the glimpsing model of speech in noise the most energetic regions of the desired speech are most important for intelligibility and thus a good predictor for intelligibility [21]. As such, it is a reasonable assumption that using only the voiced regions of the speech can yield a promising predictor for speech intelligibility.

## 4. CONCLUSION

This paper proposes a non-intrusive intelligibility metric for online processing in HAs. A clean speech signal is reconstructed by its spatio-temporal characteristics (i.e. direction of arrival and pitch) using only the noisy speech signal and utilized inside an established and reliable intrusive intelligibility metric, which requires a clean reference. The proposed non-intrusive metric has a high correlation with the original intrusive counterpart and thus is a promising method for online assessment of speech intelligibility in HAs.

## 5. REFERENCES

[1] R. W. Peters, B. C. J. Moore, and T. Baer, "Speech reception thresholds in noise with and without spectral and temporal dips for hearing-impaired and normally hearing people," *J. Acoust. Soc. Am.*, vol. 103, no. 1, pp. 577–587, 1998.

[2] J. M. Festen and R. Plomp, "Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing," *J. Acoust. Soc. Am.*, vol. 88, no. 4, pp. 1725–1736, 1990.

[3] P. C. Loizou, *Speech Enhancement: Theory and Practice*, Signal processing and communications. Taylor & Francis, 2007.

[4] Y. Hu and P. C. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," *Speech Communication*, vol. 49, no. 78, pp. 588 – 601, 2007.

[5] V. Hamacher, J. Chalupper, E. Eggers, U. Kornagel, H. Puder, and U. Rass, "Signal processing in high-end hearing aids: State of the art, challenges, and future trends," *EURASIP J. Applied Signal Process.*, vol. 18, pp. 2915–2929, 2005.

[6] T. H. Falk, V. Parsa, J. F. Santos, K. Arehart, O. Hazrati, R. Huber, J. M. Kates, and S. Scollie, "Objective quality and intelligibility prediction for users of assistive listening devices: Advantages and limitations of existing tools," *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 114–124, 2015.

[7] 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, and Language Process.*, vol. 19, no. 7, pp. 2125–2136, 2011.

[8] R. L. Goldsworthy and J. E. Greenberg, "Analysis of speech-based speech transmission index methods with implications for nonlinear operations," *J. Acoust. Soc. Am.*, vol. 116, no. 6, pp. 3679–3689, 2004.

[9] F. Chen, O. Hazrati, and P. C. Loizou, "Predicting the intelligibility of reverberant speech for cochlear implant listeners with a non-intrusive intelligibility measure," *Biomedical Signal Processing and Control*, vol. 8, no. 3, pp. 311 – 314, 2013.

[10] T. H. Falk, C. Zheng, and W.-Y. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 18, no. 7, pp. 1766–1774, 2010.

[11] M. Karbasi, A. H. Abdelaziz, and D. Kolossa, "Twin-hmm-based non-intrusive speech intelligibility prediction," in *ICASSP*, March 2016, pp. 624–628.

[12] C. Soerensen, J. B. Boldt, F. Gran, and M. G. Christensen, "Semi-non-intrusive objective intelligibility measure using spatial filtering in hearing aids," in *EUSIPCO*, August 2016, pp. 1358–1362.

[13] Y. Tang, M. Cooke, and C. Valentini-Botinhao, "Evaluating the predictions of objective intelligibility metrics for modified and synthetic speech," *Comput. Speech Lang.*, vol. 35, no. C, pp. 73–92, Jan. 2016.

[14] L. Lamarche, C. Gigure, W. Gueaieb, T. Aboulnasr, and H. Othman, "Adaptive environment classification system for hearing aids," *The Journal of the Acoustical Society of America*, vol. 127, no. 5, pp. 3124–3135, 2010.

[15] J. R. Jensen, M. G. Christensen, and S. H. Jensen, "Statistically efficient methods for pitch and DOA estimation," in *ICASSP*, May 2013, pp. 3900–3904.

[16] M. G. Christensen, P. Stoica, A. Jakobsson, and S. H. Jensen, "Multi-pitch estimation," *Signal Process.*, vol. 88, no. 4, pp. 972–983, Apr. 2008.

[17] A. Wabnitz, N. Epain, C. Jin, and A. Van Schaik, "Room acoustics simulation for multichannel microphone arrays," in *Proceedings of the International Symposium on Room Acoustics*, 2010, pp. 1–6.

[18] P. M. Djuric, "Asymptotic map criteria for model selection," *IEEE Transactions on Signal Processing*, vol. 46, no. 10, pp. 2726–2735, Oct 1998.

[19] M. Cooke, *Modelling auditory processing and organisation*, Ph.D. thesis, Cambridge University Press, 1993.

[20] 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 *Eurospeech'95. Proceedings of the 4th European Conference on Speech Communication and Speech Technology*, 18-21 September 1995, vol. 1, pp. 867–870.

[21] M. Cooke, "A glimpsing model of speech perception in noise," *The Journal of the Acoustical Society of America*, vol. 119, no. 3, pp. 1562–1573, 2006.