SOURCE SEPARATION BY FEATURE-BASED CLUSTERING OF MICROPHONES IN *AD HOC* ARRAYS

Sebastian Gergen[1], *Rainer Martin*[2], *Nilesh Madhu*[3]

[1]Bochum Institute of Technology gGmbH, Bochum, Germany
[2] Institute of Communication Acoustics, Ruhr-Universität Bochum, Germany
[3]IDLab, Department of Electronics and Information Systems, Ghent University - imec, Belgium
{*sebastian.gergen@bo-i-t.de*}, {*rainer.martin@rub.de*}, {*nilesh.madhu@ugent.be*}

ABSTRACT

Proposed is a new approach for the separation of localised sources in a reverberant environment, using *ad hoc* distributed microphones. In a first stage, the method uses a fuzzy clustering algorithm to assign microphones to individual source-dominated 'clusters'. This assignment is used to generate time-frequency masks to form an initial estimate of the source signal at each microphone of the corresponding cluster. Since microphone positions are not typically available in *ad hoc* arrays, the initial separation is exploited to derive the parameters required for delay-and-sum (DSB) beamforming. This beamforming is subsequently carried out, utilising all the microphones within a source cluster. Following this, a time-frequency post-filtering mask is estimated via the exchange of power spectra of the beamformed signals. This mask is applied to the beamformer outputs to further improve the separation. The approach is tested in three realistically-simulated rooms of different sizes, and evaluated by informal listening tests as well as by instrumental quality metrics (comprising both quality and intelligibility measures). Our experiments demonstrate that the approach can lead to a significant separation of the individual sources, yielding results very close to that obtained by DSB's where oracle knowledge of microphone and source positions is incorporated.

Index Terms— *Ad hoc* microphone array, spectral mask, source separation, beamforming, microphone clustering, IoT

1. INTRODUCTION

Multichannel source separation algorithms rely on the spatial diversity afforded by microphone arrays to accomplish their goal. Often, such algorithms are implemented in the short-time Fourier transform domain, since the spectro-temporal sparsity of speech signals allows the separation of the mixture into individual components. Several approaches exist for the source separation part, ranging from time-frequency masking approaches (e.g. [1, 2, 3]) to linear algorithms based on minimising (higher-order) statistics of the cross-power spectra (i.e. approaches based on independent component analysis (ICA), e.g. [4, 5, 6]) to algorithms that are based on a noise-canceller structure such as the generalised sidelobe canceller (GSC) (e.g. [7, 8, 9, 10]). An overview of the different separation philosophies, their advantages and trade-off's is provided in [11]. More recently, deep neural networks (DNNs) have also been used for source separation [12, 13, 14].

In all these approaches, source localisation plays an inherent part. Localisation cues are used in the mask-based approaches to apportion time-frequency points consistently to the individual source(s). For the ICA-based approaches, localisation cues are used to resolve the permutation and scaling ambiguity. In adaptive beamforming approaches, localisation cues are used to train the constructive part (usually delay-and-sum), the blocking matrix, and the noise canceller. Traditional source localisation approaches are based on microphone arrays of known geometry, synchronous signal sampling and, usually, well-calibrated microphones.

In the Internet-of-Things (IoT) era we are entering today, the idea is to enable sound capture for various applications with the use of *ad hoc* microphone arrays. Ad hoc arrays refers to any collection of microphones in a particular location. In the context of a smart home, for example, such ad hoc arrays are formed by the microphones on the smart devices in a room. The arrays can be temporarily augmented when new microphone-bearing devices are introduced into the environment (e.g. persons carrying smartphones enter the room). Such arrays thus contain a dynamic composition. Signals obtained from such arrays are neither synchronised nor is their geometry known *a priori*. Source separation in such conditions is therefore difficult.

Recent work by [15] showed that, for the case of distributed arrays, it is possible to perform *ad hoc* clustering of microphones such that each cluster of microphones is close to a source (i.e., these microphones are allocated to a source-dominated cluster). In fuzzy clustering, this clustering is soft: each microphone is allocated a fuzzy membership value (FMV) for each cluster. The clustering itself is based on predefined signal features and does not need high synchronisation accuracy between the different microphone signals

978-1-5386-8151-0/18/$31.00

(i.e. they can be *independently* sampled).

In this paper, we demonstrate how this membership value can be used, along with the assumption of disjointness of the source spectra, to separate sources captured by *ad hoc* arrays. The paper is structured as follows: first the signal model is introduced. Next, we briefly discuss the *ad hoc* clustering approach. Following this, we develop the method to extract time-frequency masks for each of the source-dominant clusters and discuss how to use this in separation frameworks. We further evaluate this idea on simulated data.

2. SIGNAL MODEL AND *AD HOC* CLUSTERING

Assume that the acoustic environment consists of N localised sources, the signals from which are captured by D microphones scattered within the environment in no particular order. The general model for the acoustic signal transmission from the N sources to a microphone d is given by:

$$x_d(t) = \sum_{n=1}^{N} \int_0^{\infty} h_{nd}(\tau) s_n(t-\tau) \mathrm{d}\tau, \tag{1}$$

with $s_n(t)$ being the n-th source signal, $h_{nd}(t)$ the impulse response from source n to microphone d, and $x_d(t)$ representing the resulting microphone signal. The received signals can be sampled, resulting in $x_d(l)$, where l is the time sample index, and then transformed to the short-time discrete Fourier domain:

$$X_d(k,b) = \mathrm{STFT}[x_d(l)], \tag{2}$$

with k and b representing the discrete frequency bin and time frame indices respectively.

In the fuzzy clustering procedure of our algorithm we utilize a feature set composed of MFCCs and their modulation spectra all of which are computed across signal segments of 4s duration. The effects of reverberation are reduced via cepstral mean normalization. For each microphone and each signal segment we obtain a feature vector $\mathbf{v}_d$ which is composed of A features, as described in more detail in [16].

Once we extracted the set of A-dimensional feature vectors $\Omega = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_D\}$ from all D *ad hoc* distributed microphones, we estimate clusters of microphones which are dominated by one of the sources in the room [17, 16]. To this end, we evaluate a least-squared error functional which is given as

$$J_m = \sum_{d=1}^{D} \sum_{n=1}^{N} (\mu_{n,d})^{\alpha} ||\mathbf{v}_d - \mathbf{u}_n||_{\beta}^2 \tag{3}$$

where $\mu_{n,d} \in [0,1]$ denotes the FMV and the distance between an estimated cluster center $\mathbf{u}_n$, $n \in \{1, ..., N\}$, and an observation $\mathbf{v}_d$ is computed as

$$||\mathbf{v}_d - \mathbf{u}_n||_{\beta}^2 = (\mathbf{v}_d - \mathbf{u}_i)^T \beta (\mathbf{v}_d - \mathbf{u}_n). \tag{4}$$

The weighting matrix β can be chosen to implement, e.g., the squared Euclidean norm, diagonal norm or Mahalanobis norm [18]. As a result of the iterative optimization process we obtain the FMV of each microphone and for any source.

3. ESTIMATION OF SPECTRAL MASKS

For the application of spectral masks, we assume that the localised sources are approximately disjoint in the short-time-frequency (T-F) plane and, therefore, only one source may be assumed to be dominant at any one T-F point (k,b). Thus, our goal is to estimate one spectral masks $\mathcal{M}_n(k,b)$ for each cluster and apply these onto the mixtures recorded at the microphones in order to get estimates of the individual, underlying source signals with a reduced amount of interference from other sources. We consider here the case of the *binary* mask given by:

$$\mathcal{M}_n(k,b) = \begin{cases} 1, & \text{if source } n \text{ is dominant at } (k,b); \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

This is the simplest separator in the T-F plane and its use is motivated further in the discussion below.

To estimate the binary masks, we propose to use the results from the fuzzy clustering of the microphones. We first identify, for each cluster n, the microphone $d = R_n$ with the highest FMV for that cluster. This microphone serves as the reference microphone for the source signal of that cluster, under the reasonable assumption that if a microphone has a high FMV for a particular cluster, the source in that cluster must dominate over the other sources for that microphone. We then compute the STFT representation $X_{R_n}(k,b)$ of this reference microphone signal of cluster n. The binary mask for cluster n is then computed as follows:

$$\mathcal{M}_n(k,b) = \begin{cases} 1 & |X_{R_n}(k,b)| > \dfrac{1}{B} \displaystyle\sum_{b-B+1}^{b} |X_{R_j}(k,b)|, \\ & j = 1, \ldots, N \text{ and } j \neq n, \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

Note that this computation of the binary mask is a generalisation of the mask traditionally used in the literature (where, typically $B = 1$). The reason behind this formulation is the following: for the case of *ad hoc* arrays, the inter-microphone distances can be quite large. Thus, for an impinging signal from a particular sound source the inter-microphone delay between the different microphones is an appreciable fraction of the frame-size used for the STFT. This can lead to a possible *jitter* in the STFT spectral amplitudes across the different microphones. Thus, if the non-averaged spectra are used for the mask generation, the masks could flip randomly due to this jitter. By averaging the spectral amplitudes across time, we reduce the effect of the jitter.

The masks are then applied onto the respective spectra $X_{i_n}(k,b)$ of all microphones i_n assigned to cluster n (a microphone is said to be assigned to cluster n if its FMV for cluster n is larger than its FMV for all other clusters):

$$\widetilde{X}_{i_n}(k,b) = X_{i_n}(k,b) \mathcal{M}_n(k,b)\,, \tag{7}$$

and, finally, we compute the inverse STFT of $\widetilde{X}_{i_n}(k,b)$ and reconstruct the time-domain signal by the overlap-add method. This yields $\widehat{s}_{i_n}$, which forms an *initial* estimate of the source signal of cluster n as received at microphone i_n. Such separation provides us with a means to now adapt the well-known beamforming algorithms to the case of *ad hoc* arrays as we discuss below.

3.1. Clustering-steered beamforming

A simple delay-and-sum beamforming can be carried out on all the microphones of a source cluster, if the relative delays between the microphones were known for that source. Since the relative locations of the microphones with respect to each other and the dominant source are unknown, one way to estimate these delays is by correlating the microphone signals with that of a *reference* microphone. However, due to the presence of the interference signal and the ambient noise, this is not directly possible. Therefore, we propose to use the initial estimates $\widehat{s}_{i_n}(l)$ for the correlation analysis. For each cluster n, we select the microphone with the highest FMV as the *reference* microphone for that cluster. Next, we use the $\widehat{s}_{i_n}(l)$ to compute the relative delay for all the microphones of the cluster with respect to the signal at the reference microphone. This is done as a *time-domain* correlation and is estimated over segments of $4s$ in length, which is also the duration across which the audio features for the fuzzy clustering are computed. Once the delays are obtained, a simple delay-and-sum beamforming (DSB) is done on the time-domain *microphone* signals $x_{i_n}(l)$. We choose to do the beamforming on the microphone signals and not on the $\widehat{s}_{i_n}(l)$, since these signals will have artefacts due to the masking, which will degrade the quality of the DSB output. Note that the periodic estimation of the delays has also the potential to compensate for the skew introduced by non-synchronous sampling at the microphones. Thus, the microphone signals within a cluster are aligned with the reference microphone.

3.2. Re-estimation of spectral masks for post-filtering

At the output of the DSB stage, we have an enhanced signal $\widehat{s}_{n,\mathrm{DSB}}(l)$ for each cluster. We can now use this enhanced signal to compute a post-filtering mask as in (6). Specifically, this post-filtering mask is given by:

$$\mathcal{M}_{n,\mathrm{DSB}}(k,b) = \begin{cases} 1 & |\widehat{S}_{n,\mathrm{DSB}}(k,b)| > \dfrac{1}{B}\displaystyle\sum_{b-B+1}^{b} |\widehat{S}_{j,\mathrm{DSB}}(k,b)|, \\ & j = 1,\ldots,N \text{ and } j \neq n, \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

The mask is then applied to $\widehat{S}_{n,\mathrm{DSB}}(k,b)$ and the time-domain signal is reconstructed, yielding the final, enhanced estimate $\widehat{s}_{n,\mathrm{DSB+Post}}$ of the source in each cluster.

4. EVALUATION & RESULTS

For the evaluation we simulate 15 microphones and two active sound sources in three different rooms (see Tab. 1). For each room, we create 10 different scenarios of source-microphone setups. In each setup, $2 \leq D_n \leq 4$ microphones for cluster $n = 1, 2$ are randomly located within the critical distance of the respective source. Additional $15 - D_1 - D_2$ microphones are placed randomly all over the room. The position of each of the sources is randomised in one or the other half of each room. We create RIRs using the method in [19]. To generate microphone signals which contain contributions from both sources we convolve male and female speech signals (clean and anechoic, TIMIT database [20]) with the respective RIRs and add the signals from both sources. Based on the microphone data we extract the audio features from signals of 4 seconds duration, sampled with 16 kHz. The spectral and cepstral analysis is carried out with a frame length of 512 samples and a frame shift of 256 samples.

In our work we use a freely available MATLAB® implementation of the fuzzy c-means algorithm [21] to estimate the clusters based on the extracted feature vectors. Main parameters for the FCM are the number of clusters which we set to $N = 2$, a weighting exponent which we select as $\alpha = 2$ and an identity matrix $\boldsymbol{\beta}$ for the distance computations in the feature space which results in the Euclidean metric. The parameter B for the time-frequency masking in (6) and (8) was set to 3, which was empirically found to be a good value.

Figure 1 demonstrates the performance of the method in terms of spectrograms. While the reference microphones, i.e. the microphones that are closest to their assigned source, contain large amounts of interference, the signals are well separated when processed with DSB+Post.

Table 1. Sizes and information about reverberation time T60 and critical distance r_H of the simulated rooms.

	Size [m^3]	T60 [ms]	r_H [m]
Room 1	$4.7 \times 3.4 \times 2.4$	340	0.6
Room 2	$6.7 \times 4.9 \times 3.5$	490	0.9
Room 3	$9.3 \times 6.9 \times 4.9$	630	1.3

For the evaluation, the enhanced signal is compared to the noisy mixture signal of the reference microphone of each cluster. The instrumental metrics used for the evaluation are the segmental SIR *improvement* seg-SIRi, the ΔPESQ i.e. the improvement in PESQ [22] and the ΔSTOI [23]. A key ingredient of our enhancement is the computation of the inter-microphone delay based on preliminary estimates of the source signal and the incorporation of this estimated delay into a DSB. To give an idea of the upper performance bound, we also present the results when using a DSB using delays computed from the true positions of the source and microphones (pDSB, or position-informed DSB). The results presented are the averaged results across all three simulated rooms and simulation scenarios. Note that for the reference

Table 2. Instrumental performance evaluation, averaged across all simulated rooms and scenarios. Performance measures for the reference signal are the absolute values and those for the enhanced signals are relative to that of the reference signal.

Method	**Cluster 1**			**Cluster 2**		
	seg-SIR (dB)	PESQ	STOI	seg-SIR (dB)	PESQ	STOI
Reference	-0.77	1.99	0.74	-1.89	1.93	0.71
	seg-SIRi (dB)	ΔPESQ	ΔSTOI	seg-SIRi (dB)	ΔPESQ	ΔSTOI
$\widehat{s}_{n,\mathrm{DSB}}(l)$	5.06	0.38	0.12	5.94	0.45	0.14
$\widehat{s}_{n,\mathrm{DSB+Post}}(l)$	5.90	0.53	0.10	6.87	0.56	0.10
$\widehat{s}_{n,\mathrm{pDSB}}(l)$	5.24	0.45	0.14	6.14	0.50	0.16

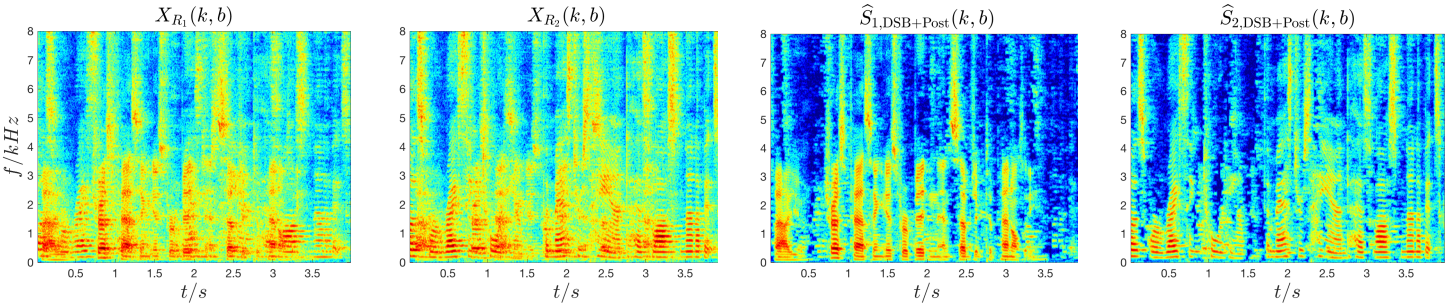

Fig. 1. Example from room 2. First two plots show the spectra of the noisy signals $x_{r_n}(l)$ at the reference microphone R_n of each cluster ($n = 1, 2$). The next two plots show the respective estimated source signals ($\widehat{s}_{n,\mathrm{DSB+Post}}(l)$).

signal, the values given are the *absolute values*. For all others, we provide the improvement relative to the reference signal. The results demonstrate that the proposed approach can yield a segmental SIRi of 6.4dB on average (DSB+Post). This result is quite in line with DSB performance for compact microphone arrays (e.g. [9, 24]). We also obtain a significant improvement in the PESQ and STOI measures for all the methods. The position-informed DSB with oracle knowledge on the positions of the sources and the microphones performs better than the DSB based on estimated time-delays. However, the performance difference is not large, which indicates that the estimated time-delays are quite close to the true values. Applying the re-estimated mask (based on the DSB outputs) further improves the PESQ and the segmental SIRi, as compared to the DSB. However, the improvement in STOI is less than for the DSB. This is understandable, since the STOI measure is based on the fidelity of the signal envelopes and the binary mask tends to distort the envelope.

5. CONCLUSION

We have presented a multi-stage approach to perform source separation using *ad hoc* microphone arrays. In the first stage, the microphones are partitioned into clusters around the localised sources. Each cluster is assumed to be dominated by a single source. The microphone with the highest fuzzy membership value in each cluster is selected as the reference microphone for that cluster. Based on the amplitude spectrum of the signal at this reference microphone, we generate a binary mask and obtain initial estimates of the source signals. These initial estimates are used to estimate the relative time-delay between all the microphones of a cluster and the reference microphone. This delay information is subsequently incorporated into a DSB framework. Based on the spectra of the DSB outputs, a binary-mask post-filter is estimated and applied to the DSB outputs to generate the final, enhanced signal.

We have evaluated this approach in a simulated scenario, in which the number of sources is assumend to be known. The estimation of the number of sources is part of ongoing research. Informal listening tests convincingly demonstrate the viability of this method for source separation using *ad hoc* arrays. The instrumental measures show a consistent and significant benefit of the different stages of the algorithm. The DSB based on estimated time-delays has a very similar performance compared to the DSB based on the (oracle) knowledge of the positions of the source and the microphones. This indicates that the time-delay computation based on initial estimates of the source signal yields an estimate very close to the true value.

The proposed approach requires only coarsely synchronised microphone signals. Further synchronisation to sampling precision is effected by time delay estimation and compensation in the DSB. The mask computation only uses the short-time power of the microphone signals, and is thus not susceptible to small synchronisation errors. These issues will be investigated in more detail in future works, in addition to tests on real recordings (e.g. with an audio scene recorded on several mobile phones).

6. REFERENCES

[1] N. Roman, D. Wang, and D. L. Brown, "Speech segregation based on sound localization," *J. Acoustical Society of America*, vol. 114, pp. 22362252, 2003.

[2] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. on Sig. Proc.*, vol. 52, no. 7, pp. 1830–1847, July 2004.

[3] D. Wang, "Time-frequency masking for speech separation and its potential for hearing aid design," *Trends in Amplification*, vol. 12, no. 4, pp. 332–353, 2008.

[4] T. Nishikawa, H. Saruwatari, and K. Shikano, "Blind source separation based on multi-stage ICA combining frequency-domain ICA and time-domain ICA," in *2002 IEEE International Conf. on Acoustics, Speech, and Signal Processing*, May 2002, vol. 1, pp. I–917–I–920.

[5] H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 5, pp. 530–538, Sept 2004.

[6] F. Nesta, P. Svaizer, and M. Omologo, "Convolutive bss of short mixtures by ica recursively regularized across frequencies," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 19, no. 3, pp. 624–639, March 2011.

[7] B. J. Yoon, I. Tashev, and A. Acero, "Robust adaptive beamforming algorithm using instantaneous direction of arrival with enhanced noise suppression capability," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, April 2007, vol. 1, pp. I–133–I–136.

[8] M. I. Mandel, R. J. Weiss, and D. P. W. Ellis, "Model-based expectation-maximization source separation and localization," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 382–394, 2010.

[9] N. Madhu and R. Martin, "A versatile framework for speaker separation using a model-based speaker localization approach," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 1900–1912, 2011.

[10] M. Taseska and E. Habets, "DOA-informed source extraction in the presence of competing talkers and background noise," *EURASIP Journal on Advances in Signal Processing*, vol. 2017, pp. 60, 12 2017.

[11] N. Madhu and A. Gückel, "Multi-channel source separation: Overview and comparison of mask- based and linear separation algorithms," in *Machine Audition: Principles, Algorithms and Systems*, W. Wang, Ed., pp. 207–245. IGI Global, USA, 2010.

[12] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 196–200.

[13] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks," *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1652–1664, Sept 2016.

[14] Yan-Hui Tu, Jun Du, Lei Sun, Feng Ma, and Chin-Hui Lee, "On design of robust deep models for chime-4 multi-channel speech recognition with multiple configurations of array microphones," in *Proc. Interspeech 2017*, 2017, pp. 394–398.

[15] S. Gergen and R. Martin, "Estimating source dominated microphone clusters in ad-hoc microphone arrays by fuzzy clustering in the feature space," in *Proc. 12th ITG Speech Communication Conference*, 2016, pp. 1–4.

[16] Sebastian Gergen, Anil Nagathil, and Rainer Martin, "Classification of reverberant audio signals using clustered ad hoc distributed microphones," *Signal Processing*, vol. 107, pp. 21–32, 2015.

[17] L.D. Zadeh, "Fuzzy sets," *Information and Control*, vol. 8, pp. 338–353, 1965.

[18] J. Bezdek, R. Ehrlich, and W. Full, "FCM: The fuzzy c-means clustering algorithm," *Computers & Geosciences*, vol. 10, no. 2-3, pp. 191–203, 1984.

[19] S. Gergen, C. Borß, N. Madhu, and R. Martin, "An optimized parametric model for the simulation of reverberant microphone signals," in *Proc. of the International Conference on Signal Processing, Communications and Computing (ICSPCC 2012)*, 2012.

[20] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "Timit acoustic-phonetic continuous speech corpus," 1993, Linguistic Data Consortium, Philadelphia.

[21] J. Abonyi, *Fuzzy Clustering and Data Analysis Toolbox*, April 2005.

[22] *Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs*, ITU-T Recommendation P.862, 2001.

[23] C.H. Taal, R.C. Hendriks, R. Heusdens, and J. Jensen, "An Algorithm for Intelligibility Prediction of TimeFrequency Weighted Noisy Speech," *TASL*, vol. 19, no. 7, pp. 2125 – 2136, 2011.

[24] P. Vary and R. Martin, *Digital Speech Transmission*, Wiley, 2006.