Bidirectional Long-Short Term Memory Network-based Estimation of Reliable Spectral Component Locations

Aaron Nicolson and Kuldip K. Paliwal

Signal Processing Laboratory, Griffith University, Brisbane, Australia

aaron.nicolson@griffithuni.edu.au, k.paliwal@griffith.edu.au

Abstract

An accurate Ideal Binary Mask (IBM) estimate is essential for Missing Feature Theory (MFT)-based speaker identification, as incorrectly labelled spectral components (where a component is either reliable or unreliable) will degrade the performance of an Automatic Speaker Identification (ASI) system adversely in the presence of noise. In this work a Bidirectional Recurrent Neural Network (BRNN) with Long-Short Term Memory (LSTM) cells is proposed for improved IBM estimation. The proposed system had an average IBM estimate accuracy improvement of 4.5% and an average MFT-based speaker identification accuracy improvement of 3.1% over all tested SNR_{dB} levels, when compared to the previously proposed Multilayer Perceptron (MLP)-IBM estimator. When used for speech enhancement the proposed system had an average MOS-LQO (objective quality measure) improvement of 0.32 and an average QSTI (objective intelligibility measure) improvement of 0.01 over all tested SNR_{dB} levels, when compared to the MLP-IBM estimator. The results presented in this work highlight the effectiveness of the proposed BRNN-IBM estimator for MFT-based speaker identification and IBM-based speech enhancement.

Implementation and Availability: The proposed BRNN-IBM estimator and further results are available at `https://github.com/anicolson/bidirectional_2018`

Index Terms: ideal binary mask estimation, missing feature theory, robust speaker identification, speech enhancement

1. Introduction

Speech is often not the only sound source present in real-world environments, making the problem of speaker identification more difficult. An Automatic Speaker Identification (ASI) system should not be affected by non-target sources, or in other words, the system must be robust to noise. Missing Feature Theory (MFT) is a technique used to negate the effects of non-target sources during classification. Many MFT-based methods were proposed when Gaussian Mixture Model - Hidden Markov Model (GMM-HMM) Automatic Speech Recognition (ASR) systems were prominent, providing a significant increase to their robustness [1][2].

MFT is based upon the human perception system and its ability to perform auditory scene analysis [3]. The psychological process involves segregating components that come from different sound sources, and grouping components that come from the same sound source. A component may have a mixture of multiple sources present, requiring a criterion to decide if a component reliably represents the target speech. In MFT, a component is classified as either a reliable or an unreliable representation of the target speech. The impact of non-target sources on recognition performance is reduced by treating the unreliable components as 'missing'. Cooke et al. [1] described that the solution to the following two problems is required for MFT-based speech recognition:

1. The identification of reliable spectral components.
2. The modification of recognition algorithms to handle incomplete data.

The work presented in this paper is focused on the first problem.

A binary mask is able to identify the time-frequency locations of reliable components [4]. When finding a binary mask, clean speech is treated as the target source, and non-target sources are treated as noise. Noisy speech is a mixture of both clean speech and noise. An Ideal Binary Mask (IBM) is computed from the spectral components of both the clean and noisy speech, where reliable components have an SNR above a set threshold [4]. However, in real-world environments only the noisy speech components are observed. This requires the IBM to be estimated from the noisy speech components, a task that Raj et al. [5] described as the most difficult aspect of MFT. An MFT-based ASI system will suffer if reliable and unreliable spectral components are misclassified. This means that an accurate IBM estimator is essential for an MFT-based ASI system.

An early IBM estimation approach used the Gaussian distribution of the noise spectrum to estimate the original speech via Spectral Subtraction (GSS) [6]. A Multilayer Perceptron (MLP) was recently used to estimate the IBM for an MFT-based ASI system [7]. MLP-IBM estimators have also been used for tasks other than MFT; an MLP-IBM estimator with an inverse fast Fourier transform layer was used for source separation [8]. Related source separation tasks have used Recurrent Neural Networks (RNN) with Long-Short Term Memory (LSTM) cells to produce state-of-the-art results [9][10].

This work aims to solve the first problem of MFT proposed by Cooke et al. [1] by using a Bidirectional RNN (BRNN) [11] with LSTM cells for IBM estimation (Section 2). The performance of RNNs for speech separation indicates the potential performance of a BRNN-IBM estimator for MFT. By producing an accurate IBM estimator, it is hoped that future research into MFT-based methods for modern ASR systems is encouraged. The proposed BRNN-IBM estimator and previous IBM estimators are compared in terms of their IBM estimation accuracy (Section 4) and their MFT-based speaker identification accuracy (Subsection 5.1). While the emphasis of this work is to evaluate the proposed BRNN-IBM estimator for MFT-based ASI, its speech enhancement performance is also evaluated (Subsection 5.2). Conclusions are drawn in Section 6.

2. BRNN-IBM Estimator

The proposed system shown in Figure 1 takes as its input the noisy speech spectral components, $\mathbf{X}(n)$, of frame n. The system then estimates the IBM of $\mathbf{X}(n)$, as given by $\hat{\mathbf{y}}(n)$. The spectral feature types used in this work include Magnitude

Spectrum (MS) components and Log-Spectral Subband Energy (LSSE) components[1]. 'BRNN-MS-IBM estimator' denotes the BRNN used to estimate the IBM for the noisy speech MS components and 'BRNN-LSSE-IBM estimator' denotes the BRNN used to estimate the IBM for the noisy speech LSSE components.

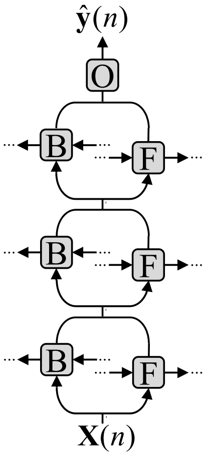

Figure 1: *Proposed BRNN-IBM estimator. The input features $\mathbf{X}(n)$ are the noisy speech MS/LSSE components for the n^{th} frame. The output of the network $\hat{\mathbf{y}}(n)$ is the IBM estimate for the n^{th} frame.*

The IBM is used as the training target for a BRNN-IBM estimator. To compute the IBM of a noisy speech MS/LSSE component, the SNR$_{\text{dB}}$ of each component is found. When clean speech is mixed with uncorrelated additive noise, the k^{th} noisy speech MS/LSSE component, $\mathrm{X}(n,k)$ for the n^{th} frame can be modeled as the sum of its corresponding clean speech MS/LSSE component $\mathrm{S}(n,k)$ and noise MS/LSSE component $\mathrm{D}(n,k)$:

$$\mathrm{X}(n,k) = \mathrm{S}(n,k) + \mathrm{D}(n,k). \tag{1}$$

The SNR$_{\text{dB}}$ for the k^{th} noisy speech MS/LSSE component of frame n is calculated by

$$\mathrm{SNR}_{\mathrm{dB}}(n,k) = 20 * \log_{10}\left(\frac{\mathrm{S}(n,k)}{\mathrm{X}(n,k) - \mathrm{S}(n,k)}\right). \tag{2}$$

It is assumed that noisy speech MS/LSSE components with an SNR$_{\text{dB}}$ above a set threshold θ are reliable estimates of the corresponding clean speech MS/LSSE components. The IBM is found by

$$\mathrm{IBM}(n,k) = \begin{cases} 1, & \text{if } \mathrm{SNR}_{\mathrm{dB}}(n,k) > \theta \\ 0, & \text{otherwise,} \end{cases} \tag{3}$$

where $\theta = 0$ is used in this work [4].

The proposed BRNN-IBM estimator in Figure 1 consists of three BRNN layers, with LSTM cells as in [13]. Each hidden layer has a forward LSTM cell, **F**, with 512 units and a backward LSTM cell, **B**, with 512 units. The output layer, **O**, is a sigmoidal fully-connected layer with its number of units equal to the number of input dimensions. BRNNs with LSTM cells are better able to model time sequences than MLPs, due to the internal memory of the LSTM cells. The proposed BRNN-IBM estimator is able to store information about the noise sources and the target speech over time, enabling it to make more informed decisions about a component's reliability.

3. Datasets and Experiment Setup

The TIMIT corpus [14] (16 kHz, single-channel) which consists of 630 speakers with 10 utterances each, was used as the clean speech set. For the LSSE-IBM estimators and the speaker identification tests, it was required that the speakers in the training and test sets were matched. Therefore, the si^* and sx^* subsets were used for training (5040 utterances) and the sa^* subset was used for testing (1260 utterances). For the MS-IBM estimators and the speech enhancement tests, it was required that the speakers in the training and the test sets were separate. The si^* and sx^* subsets were split into 462 speakers for training (3696 utterances) and 168 speakers for testing (1344 utterances). The sa^* subset was removed as the utterances are the same across all speakers.

The RSG-10 noise dataset [15] (16 kHz, single-channel), which includes 24 different noise sources as described in Table 1, was used as the noise set. Noise was added to the speech at an SNR$_{\text{dB}}$ level of 0 to 30 dB, in 5 dB increments. The entire test set was used at each SNR$_{\text{dB}}$ level for each experiment.

Table 1: *The 24 noise sources included in the RSG-10 noise dataset [15]. f_c is the cutoff frequency.*

Source	**MM:SS**	**Description**
Sinusoid	00:57	1000 Hz
Pink noise	02:59	Equal energy per 1/3-oct
White noise	03:54	Equal energy per Hz
White -6 dB/oct	03:54	$f_c = 250$ Hz, -6 dB/oct
White -12 dB/oct	03:54	$f_c = 250$ Hz, -12 dB/oct
Speech noise	03:54	Average speech spectrum
M 109	03:54	30 km/h
Buccaneer	03:54	Pilot 190 Knots 1000 Feet
Leopard 2	03:54	70 km/h
Wheel carrier	03:54	50-60 km/h
Buccaneer	03:54	450 Knots 300 Feet
Lynx	03:54	Platform
Leopard 1	03:54	70 km/h
Operations room	03:54	Destroyer operations room
Destroyer	03:54	Engine room
Machine gun	03:54	Calibre 0.50 repeated
HF radio	03:54	Noise from HF radio channel
STITEL	03:54	STI test signal
Voice babble	03:54	Canteen, 100 people
F-16 two-seat	03:54	300-600 Feet, 500 Knots
Car Factory	03:54	Electrical welding
Car Factory	03:54	Car production hall
Car Volvo-340	03:54	120 km/h, asphalt road
Car Volvo-340	03:54	50 km/h, brick road

A frame length of 32 ms and a shift of 16 ms was used for signal framing. Features were computed from the 512-point Discrete Fourier Transform (DFT) of the frames. The 257-point single-sided MS included both the DC frequency component and the Nyquist frequency component. The LSSE components were computed from a 26 filter mel-scaled filterbank as in [7].

Each neural network employed the following architecture and training strategy:

- A fully-connected output layer with sigmoidal units.
- Cross-entropy as the loss function.

[1] LSSE components are also known as log-filterbank energy components [12].

- The *Adam* algorithm [16] for gradient descent optimisation.
- 5% of the training set was used as a validation set.
- Clean speech signals, noise signals, and SNR$_{\text{dB}}$ levels were all randomly chosen for each mini-batch.
- A random section of each noise signal was extracted for the mini-batch.
- A mini-batch size of 20 noisy speech signals.
- Validation error was found every 50 mini-batches, and the network parameters saved if the lowest validation error was achieved.
- The network parameters were replaced by the saved network parameters every 2 000 mini-batches.
- A total of 100 000 mini-batches.

4. IBM Estimate Accuracy

An IBM estimator must be able to correctly classify the time-frequency components of noisy speech into reliable and unreliable components. The binary variable $\alpha(n,k)$ determines if the k^{th} component of the n^{th} frame has been correctly classified:

$$\alpha(n,k) = \begin{cases} 1 & \text{if } \hat{y}(n,k) = \text{IBM}(n,k) \\ 0 & \text{if } \hat{y}(n,k) \neq \text{IBM}(n,k), \end{cases} \tag{4}$$

where $\hat{y}(n,k)$ is the IBM estimate of the noisy speech MS/LSSE component, and IBM(n,k) is its target value. The accuracy of the IBM estimator is found over all N frames in the test set:

$$Accuracy\ (\%) = 100 \times \frac{\sum_n \sum_k \alpha(n,k)}{NK}, \tag{5}$$

where K is the total number of components in a frame.

LSSE-IBM estimators were compared by their LSSE-IBM estimate accuracy, as shown in Table 2. The proposed BRNN-LSSE-IBM estimator (3 hidden layers, 1024 units per layer) was compared to an MLP-LSSE-IBM estimator (5 hidden layers, 1024 Rectified Linear Units (ReLU) [17] per layer) and the GSS-LSSE-IBM estimator. The proposed BRNN-LSSE-IBM estimator was the most accurate at all SNR$_{\text{dB}}$ levels, with an average improvement of 4.5% over all tested SNR$_{\text{dB}}$ levels when compared to the MLP-LSSE-IBM estimator.

Table 2: *LSSE-IBM estimate accuracy (%). The highest accuracy at each SNR$_{dB}$ level is shown in boldface.*

	SNR Level (dB)						
LSSE-IBM Estim.	**0**	**5**	**10**	**15**	**20**	**25**	**30**
GSS-LSSE-IBM [6]	70.5	62.8	54.9	47.2	40.5	35.1	31.0
MLP-LSSE-IBM [7]	91.2	90.3	89.6	89.7	90.6	91.7	92.6
BRNN-LSSE-IBM	**94.6**	**94.4**	**94.5**	**94.9**	**95.6**	**96.2**	**96.8**

MS-IBM estimators were also compared by their MS-IBM estimate accuracy, as shown in Table 3. The proposed BRNN-MS-IBM estimator (3 hidden layers, 1024 units per layer) was compared to an MLP-MS-IBM estimator (5 hidden layers, 1024 ReLUs per layer) and the GSS-MS-IBM estimator. The proposed BRNN-MS-IBM estimator was the most accurate at all SNR$_{\text{dB}}$ levels, with an average improvement of 3.1% over all tested SNR$_{\text{dB}}$ levels when compared to the MLP-MS-IBM estimator.

Table 3: *MS-IBM estimate accuracy (%). The highest accuracy at each SNR$_{dB}$ level is shown in boldface.*

	SNR Level (dB)						
MS-IBM Estim.	**0**	**5**	**10**	**15**	**20**	**25**	**30**
GSS-MS-IBM [6]	78.8	73.8	67.9	61.4	54.9	49.0	44.1
MLP-MS-IBM [7]	87.9	86.5	85.4	84.9	85.4	86.7	88.4
BRNN-MS-IBM	**89.8**	**88.8**	**88.3**	**88.3**	**89.1**	**90.4**	**92.0**

5. IBM Estimate Applications

5.1. Automatic Speaker Identification

Table 4: *Marginalisation vs. no marginalisation speaker identification accuracy (%) with an SSC-GMM-ASI system – clean speech is given to compute the LSSE-IBM. The accuracy of the system on clean speech is 96.8%. The highest accuracy at each SNR$_{dB}$ level is shown in boldface.*

	SNR Level (dB)						
Marg.	**0**	**5**	**10**	**15**	**20**	**25**	**30**
No [7]	8.9	14.1	27.5	48.0	69.7	84.8	92.5
Yes [1][7]	**59.1**	**71.6**	**78.7**	**85.1**	**89.6**	**92.6**	**94.2**

An MFT-based ASI system was used to compare the LSSE-IBM estimators. The Spectral Subband Centroid (SSC) [18][19] diagonal-covariance GMM-ASI system from [7] was used to compare the LSSE-IBM estimators, where marginalisation [1] is the MFT-based method used by the system. Marginalisation is a classifier-compensation method that ignores the unreliable components during classification. The ASI system has 32-mixture GMM speaker models, and uses an LSSE-IBM estimate to identify the reliable SSC components[2]. The performance difference of the SSC-GMM-ASI system when marginalisation was used can be seen in Table 4, where the system became significantly more robust at all SNR$_{\text{dB}}$ levels. An LSSE-IBM was used to obtain the marginalisation results in Table 4 (i.e. clean and noisy speech was used to compute the LSSE-IBM).

Table 5: *Speaker identification accuracy (%) with a marginalisation-based SSC-GMM-ASI system. LSSE-IBM estimators were used to identify the reliable SSCs. The highest accuracy at each SNR$_{dB}$ level is shown in boldface*

	SNR Level (dB)						
LSSE-IBM Est.	**0**	**5**	**10**	**15**	**20**	**25**	**30**
GSS-LSSE-IBM [6]	37.2	54.1	67.6	78.3	86.0	90.6	93.7
MLP-LSSE-IBM [7]	51.2	63.4	72.4	79.3	86.0	91.1	**94.1**
BRNN-LSSE-IBM	**55.6**	**69.4**	**77.2**	**82.9**	**88.2**	**91.8**	**94.1**

An LSSE-IBM estimate is required by the ASI system when clean speech is not given to compute the LSSE-IBM. Table 5 shows the accuracy of the ASI system when the LSSE-IBM estimators were used. The proposed BRNN-LSSE-IBM estimator achieved the best speaker identification accuracy at all SNR$_{\text{dB}}$ levels, with an average improvement of 3.1% over all tested

[2] SSC components are computed from the same 26 filter mel-scale filterbank as the LSSE components.

SNR$_{\mathrm{dB}}$ levels when compared to the MLP-LSSE-IBM estimator. The results of the proposed BRNN-LSSE-IBM estimator more closely match the LSSE-IBM results in Table 4 than the results of the MLP-LSSE-IBM estimator.

5.2. Speech Enhancement

The proposed BRNN-MS-IBM estimator can be used for speech enhancement by using the MS-IBM estimate as a gain function [20]. The unreliable components of the noisy speech MS are suppressed completely. Figure 2 displays the results of using the proposed BRNN-MS-IBM estimator for speech enhancement.

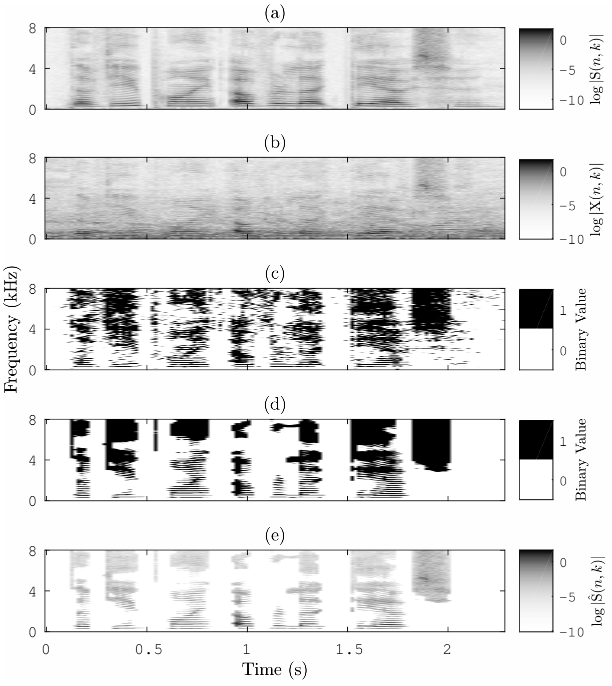

Figure 2: *(a) Clean speech magnitude spectrogram of a female uttering: "The viewpoint overlooked the ocean". (b) Noisy speech magnitude spectrogram ('operations room' at 0 dB). (c) MS-IBM for (b). (d) MS-IBM estimate for (b) using the proposed BRNN-MS-IBM estimator. (e) Resultant enhanced speech magnitude spectrogram after (d) has been applied.*

Objective measures were used to evaluate the quality and intelligibility of the enhanced speech produced by the proposed BRNN-MS-IBM estimator. The proposed BRNN-MS-IBM estimator was compared to the MLP-MS-IBM estimator, the Minimum Mean Square Error - Log-Spectral Amplitude (MMSE-LSA) estimator [21], the perceptually Motivated Bayesian MS Estimator (pMMSE) [22] with noise estimation from [23], and noisy speech. Mean Opinion Score - Listening Quality Objective (MOS-LQO) (P.800.1) [24] was used for objective quality evaluation, where Wideband Perceptual Evaluation of Speech Quality (Wideband PESQ) (P.862.2) [25] was used to obtain the MOS-LQO. Table 6 shows the average MOS-LQO over the test set. It can be seen that the proposed BRNN-MS-IBM estimator achieved the highest average MOS-LQO at most SNR$_{\mathrm{dB}}$ levels (0, 15, 20, 25, and 30 dB), with an average improvement of 0.32 over all tested SNR$_{\mathrm{dB}}$ levels when compared to the MLP-MS-IBM estimator.

Table 6: *Average MOS-LQO for the speech enhancement methods (obtained using Wideband PESQ). The highest average MOS-LQO at each SNR$_{dB}$ level is shown in boldface. Clean speech is given to compute the MS-IBM.*

	SNR Level (dB)						
Method	**0**	**5**	**10**	**15**	**20**	**25**	**30**
Noisy speech	1.17	1.31	1.56	1.93	2.46	3.04	3.56
MLP-MS-IBM [7]	1.42	1.60	1.91	2.35	2.85	3.31	3.65
MMSE-LSA [21]	1.43	1.72	2.11	2.57	3.05	3.50	3.86
pMMSE [22] [23]	1.53	**1.85**	**2.23**	2.64	3.05	3.42	3.71
BRNN-MS-IBM	**1.58**	1.82	2.20	**2.71**	**3.26**	**3.72**	**4.06**
MS-IBM	2.10	2.50	2.96	3.42	3.82	4.12	4.33

The Quasi-stationary Speech Transmission Index (QSTI) was used for objective intelligibility testing, and is more correlated with subjective intelligibility testing than the Speech Transmission Index (STI) [26]. Table 7 shows the average QSTI over the test set. The proposed BRNN-MS-IBM estimator was able to score the highest average QSTI at all SNR$_{\mathrm{dB}}$ levels, with an average improvement of 0.01 over all tested SNR$_{\mathrm{dB}}$ levels when compared to the MLP-MS-IBM estimator. The average MOS-LQO and QSTI of the MS-IBM are also shown in Table 6 and 7, respectively, indicating the performance upper limit of the proposed BRNN-MS-IBM estimator for speech enhancement.

Table 7: *Average QSTI for the speech enhancement methods. The highest average QSTI at each SNR$_{dB}$ level is shown in boldface. Clean speech is given to compute the MS-IBM.*

	SNR Level (dB)						
Method	**0**	**5**	**10**	**15**	**20**	**25**	**30**
pMMSE [22] [23]	0.82	0.87	0.91	0.94	0.96	0.98	0.99
MMSE-LSA [21]	0.83	0.88	0.92	0.95	0.97	0.98	0.99
Noisy speech	0.84	0.89	0.93	**0.96**	**0.98**	**0.99**	**1.00**
MLP-MS-IBM [7]	0.85	0.90	0.93	**0.96**	**0.98**	**0.99**	0.99
BRNN-MS-IBM	**0.88**	**0.91**	**0.94**	**0.96**	**0.98**	**0.99**	**1.00**
MS-IBM	0.90	0.92	0.95	0.97	0.98	0.99	1.00

6. Conclusion

The robustness of an MFT-based ASI system suffers when given a poor IBM estimate. In this work, a BRNN with LSTM cells was used for IBM estimation. The proposed BRNN-LSSE-IBM had an average IBM estimate accuracy improvement of 4.5% and an average MFT-based speaker identification accuracy improvement of 3.1% over all tested SNR$_{\mathrm{dB}}$ levels when compared to the previously proposed MLP-LSSE-IBM estimator. When used for speech enhancement, the proposed BRNN-MS-IBM estimator had an average MOS-LQO improvement of 0.32 and average QSTI improvement of 0.01 over all tested SNR$_{\mathrm{dB}}$ levels when compared to the MLP-MS-IBM estimator. The results of the proposed method have demonstrated that it is the most accurate IBM estimator for MFT-based ASI systems and IBM-based speech enhancement. By providing an accurate IBM estimator, it is hoped that research into MFT-based methods for modern ASR systems is encouraged. This work will be used in the development of future MFT-based methods for modern ASR systems.

7. References

[1] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech communication*, vol. 34, no. 3, pp. 267–285, 2001.

[2] B. Raj, M. L. Seltzer, and R. M. Stern, "Reconstruction of missing features for robust speech recognition," *Speech communication*, vol. 43, no. 4, pp. 275–296, 2004.

[3] A. S. Bregman, *Auditory scene analysis: The perceptual organization of sound*. MIT press, 1994.

[4] D. Wang, *On Ideal Binary Mask As the Computational Goal of Auditory Scene Analysis – Speech Separation by Humans and Machines*. Boston, MA: Springer US, 2005, pp. 181–197.

[5] B. Raj and R. M. Stern, "Missing-feature approaches in speech recognition," *IEEE Signal Processing Magazine*, vol. 22, no. 5, pp. 101–116, 2005.

[6] P. Renevey and A. Drygajlo, "Missing feature theory and probabilistic estimation of clean speech components for robust speech recognition," in *Sixth European Conference on Speech Communication and Technology*, 1999, pp. 2627–2630.

[7] A. Nicolson, J. Hanson, J. Lyons, and K. Paliwal, "Spectral subband centroids for robust speaker identification using marginalization-based missing feature theory," *International Journal of Signal Processing Systems*, vol. 6, no. 1, pp. 12–16, 2018.

[8] Y. Wang and D. Wang, "A deep neural network for time-domain signal reconstruction," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 4390–4394.

[9] P. S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Joint optimization of masks and deep recurrent neural networks for monaural source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 12, pp. 2136–2147, 2015.

[10] Z. Chen, S. Watanabe, H. Erdogan, and J. R. Hershey, "Speech enhancement and recognition using multi-task learning of long short-term memory recurrent neural networks," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015, pp. 3274–3278.

[11] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[12] K. K. Paliwal, "Decorrelated and liftered filter-bank energies for robust speech recognition," in *Proc. 6th European Conf. Speech Communication and Technology, EUROSPEECH-99*, vol. 2, 1999, pp. 85–88.

[13] F. A. Gers and J. Schmidhuber, "Recurrent nets that time and count," in *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium*, vol. 3. IEEE, 2000, pp. 189–194.

[14] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM. NIST speech disc 1-1.1," *NASA STI/Recon Technical Report N*, vol. 93, Feb. 1993.

[15] H. J. Steeneken and F. W. Geurtsen, "Description of the RSG-10 noise database," *Report IZF 1988-3, TNO Institute for Perception, Soesterberg, The Netherlands*, 1988.

[16] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014.

[17] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th International Conference on International Conference on Machine Learning*, 2010, pp. 807–814.

[18] K. K. Paliwal, "Spectral subband centroid features for speech recognition," in *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, vol. 2, 1998, pp. 617–620.

[19] B. Gajic and K. K. Paliwal, "Robust feature extraction using subband spectral centroid histograms," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings*, vol. 1, 2001, pp. 85–88.

[20] 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.

[21] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE transactions on acoustics, speech, and signal processing*, vol. 33, no. 2, pp. 443–445, 1985.

[22] P. C. Loizou, "Speech enhancement based on perceptually motivated bayesian estimators of the magnitude spectrum," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 857–869, 2005.

[23] T. Gerkmann and R. C. Hendriks, "Unbiased mmse-based noise power estimation with low complexity and low tracking delay," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1383–1393, 2012.

[24] ITU-T Recommendation P.800.1, "Mean opinion score (MOS) terminology," 2006.

[25] ITU-T Recommendation P.862.3, "Application guide for objective quality measurement based on recommendations P.862, P.862.1 and P.862.2," 2007.

[26] B. Schwerin and K. Paliwal, "An improved speech transmission index for intelligibility prediction," *Speech Communication*, vol. 65, pp. 9 – 19, 2014.