INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Monaural Source Separation Using a Random Forest Classifier

Cosimo Riday, Saurabh Bhargava*, Richard H.R. Hahnloser, and Shih-Chii Liu*

Institute of Neuroinformatics, University of Zürich and ETH Zürich, Switzerland

ridayc, saurabh, rich, shih@ini.ethz.ch

ABSTRACT

We address the problem of separating two audio sources from a single channel mixture recording. A novel method called Multi Layered Random Forest (MLRF) that learns a binary mask for both the sources is presented. Random Forest (RF) classifiers are trained for each frequency band of a source spectrogram. A specialized set of linear transformations are applied to a local time-frequency (T-F) neighborhood of the mixture that captures relevant local statistics. A sampling method is presented that efficiently samples T-F training bins in each frequency band. We draw equal numbers of dominant (more power) training samples from the two sources for RF classifiers that estimate the Ideal Binary Mask (IBM). An estimated IBM in a given layer is used to train a RF classifier in the next higher layer of the MLRF hierarchy. On average, MLRF performs better than deep Recurrent Neural Networks (RNNs) and Non-Negative Sparse Coding (NNSC) in signal-to-noise ratio (SNR) of reconstructed audio, overall T-F bin classification accuracy, as well as PESQ and STOI scores. Additionally, we demonstrate the ability of the MLRF to correctly reconstruct T-F bins of the target even when the latter has lower power in that frequency band.

Index Terms: monaural source separation, random forest, deep learning, CASA.

1. Introduction

Approaches that address the problem of estimating the underlying source signals of a mixture [1] can be extremely useful for speech recognition systems [2], in medical signal processing [3], [4]; as well as in hearing aid devices [5], [6]. The problem becomes more challenging when only monaural mixture recordings are available [7] and therefore is still an active area of research.

One approach known as Computational Auditory Scene Analysis (CASA) uses binary masking for demixing [8], [9]. One goal of CASA [10] is to estimate the Ideal Binary Mask (IBM) consisting of the set of T-F bins associated with a given source. It has been shown that under certain conditions the reconstructed speech using IBMs provides the optimal SNR gain among all binary masks (BMs) [11] and is highly intelligible [12]. Therefore the T-F bins of IBMs are useful as training labels for classification systems such as commonly implemented deep learning architectures [13]–[15] or the Multi Layered Random Forest method proposed in this paper.

Other approaches such as speech enhancement techniques include spectral subtraction, Wiener filtering, and subspace based methods [16]. The model-based approaches such as Non-Negative Sparse Coding (NNSC) [17], [18] have been extensively used for separating audio signals such as speech, noise and music. Recently deep learning architectures such as Recurrent Neural Networks (RNNs) have shown state-of-the-art performance in monaural source separation (MSS) [13], [19].

We propose a novel CASA method, called Multi Layered Random Forest (MLRF) for learning BMs for both audio sources in a mixture. The Random Forest (RF) method maps a high-dimensional input vector (features) to an output class (in our case 0 or 1) which is (a value of) a T-F bin in the BM. After a first classification, the BM is used to create additional features that are provided to the subsequent RF classifiers (later called layers). The final BM is then applied to a test mixture spectrogram to obtain estimates of the source spectrograms that are inverted back to the audio domain.

In summary, our work makes four contributions. First, we train RF classifiers for individual frequency bands similar to [15], thereby enhancing the performance of MSS. This is motivated by the observation that T-F neighborhood statistics of structured audio data such as speech, music and others are frequency dependent [20]. We generate training mixtures by linearly adding T-F bins of the training spectrograms of the two sources, forming a training mixture spectrogram. The IBM determines the dominant source in a given T-F bin. Second, we introduce a sampling method that returns the temporal coordinates of both training spectrograms required to make a source specific training mixture sample for a given frequency band. To achieve a high classification recall for both sources simultaneously, an equal number of dominant samples are chosen for each source. Third, the input feature vectors are constructed using various linear transformations on a frequency specific T-F neighborhood. Four, we compare our results to a state-of-the-art deep learning approach [13], to NNSC [17] using generalized Kullback-Leibler (KL) divergence criteria and its BM version NNSC(BM) and demonstrate significant improvements. The MLRF method is described in Section 2, the experimental setup in Section 3, the results are discussed in Section 4, and conclusions are presented in Section 5.

2. Methods

2.1. Dynamic range compression

To provide dynamic range compression, input training and testing signals are locally normalized using a Gaussian window:

$$\tilde{s}(t) = \frac{s(t)}{\sqrt{\mathcal{N}(t,\mu,\sigma) * s^2(t)}} \quad (1)$$

where $\mathcal{N}(t,\mu,\sigma) = \sqrt{\frac{1}{2\pi\sigma^2}} \cdot e^{\frac{-(t-\mu)^2}{2\sigma^2}}$ is the normalized Gaussian function with mean μ and variance σ^2, $s(t)$ is the original audio signal, $\tilde{s}(t)$ the normalized signal, and * the convolution

*equal contribution to this paper

http://dx.doi.org/10.21437/Interspeech.2016-252

operator. For large Gaussian half-widths, this is comparable to root mean squared (RMS) [21] normalization of the whole signal.

2.2. Spectrograms and binary masks

As many separation tasks become more tractable after being projected to a higher dimensional representation, we create the spectrogram S of a normalized audio signal $\tilde{s}$ by computing the absolute value of its Short Time Fourier Transforms (STFT):

$$S = |\mathrm{STFT}(\tilde{s})| \tag{2}$$

During the training phase, the two independent training audio sources are processed using (1) and (2) to obtain the respective training spectrograms X_{tr} and Y_{tr}. T-F bins of an IBM are computed by comparing these source spectrograms for the same frequency f and different times τ_1 and τ_2:

$$IBM_{tr}[(\tau_1,\tau_2),f] = \begin{Bmatrix} 1 & X_{tr}[\tau_1,f] \geq Y_{tr}[\tau_2,f] \\ 0 & X_{tr}[\tau_1,f] < Y_{tr}[\tau_2,f] \end{Bmatrix} \tag{3}$$

In the literature, $\tau_1 = \tau_2$, whereas the above definition allows for comparison of time-shifted spectrograms used in training (see Section 2.4).

During the test phase, the two independent test audio sources are processed using (1) and added to obtain an artificial test audio mixture. This mixture is processed using (2) to obtain the test mixture spectrogram Z_T.

2.3. Frequency-wise classification with random forests

The IBM of mixture data can be estimated with a binary classifier trained on synthesized mixtures. We train individual classifiers on single frequency bands to predict the IBM labels 0 and 1.

We choose RFs [22] as classifiers, motivated by their following properties: (1) Invariance towards monotonic transformations of input dimensions (i.e. exponentiation, or log transformation of spectrogram values); (2) robustness towards irrelevant input dimensions; (3) only a small set of dependency parameters are used (in particular the number of randomly chosen dimensions for splitting at each decision node); (4) trivial parallelization of the algorithm during training and prediction; (5) good interpretability of the separation task based on decision tree structures.

2.4. Sampling the training T-F bins

Mixtures created from the addition of diverse time-shifted versions of the training set are converted to training mixture spectrograms. From these mixtures, we obtain a single IBM label (0 or 1) for each T-F bin of a source-specific IBM. If L_x and L_y are the number of time bins of the training spectrograms, this results in $L_x \cdot L_y$ potential mixtures. To provide simultaneous high recall for both sources, we take an equal number of samples from both IBM labels.

To randomly sample for a specific label and obtain time points τ_1 and τ_2 for both sources, we pre-sort both sources at the particular frequency according to their T-F bin power and find the fraction of bins in source Y^f (Y at frequency f) that have less power than a particular bin in source X^f and vice versa. Drawing one sample from the resulting cumulative distribution requires a binary search on X^f and Y^f and therefore O(log(L_x) +log(L_y)) operations.

2.5. Feature extraction

The RF algorithm learns decision trees on input feature vectors comprising three subsets: The first subset consists of the local neighborhood (of dimension $\omega_\tau \cdot \omega_f$, where ω_τ and ω_f are the temporal and spectral window widths) surrounding each T-F bin. An example is shown for the sample location τ_0, f_0 (see Figure 1, top).

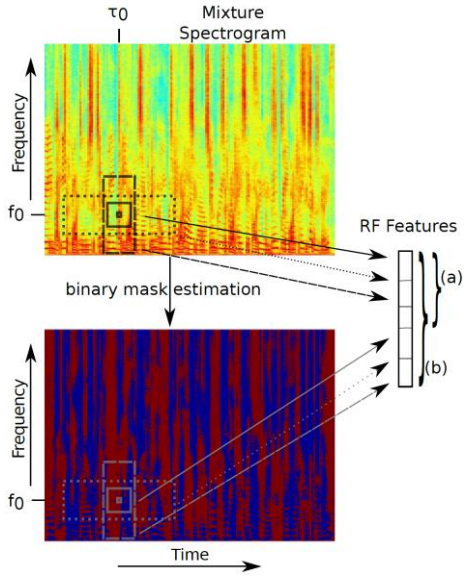

Figure 1: *Top: Multiple neighborhood windows (dotted and dashed lines) around a T-F bin* $(\boldsymbol{\tau_0, f_0})$ *in a log-power spectrogram of a speech mixture sample. Bottom: BM estimate in the first MLRF layer with blue color coding as 0 and red as 1. Input vectors to the first layer (a) are formed from linear combinations of windowed inputs from the mixture spectrogram and to the second layer (b) are formed from linear combinations of windowed inputs from the mixture spectrogram and from the first-layer mask estimates (slanted arrows).*

For the second subset, we quantify the dissimilarities within the T-F neighborhood between the set z_0 of mixture samples with only label 0, and the set $z_{0,1}$ with an equal number of both labels. A measure of this dissimilarity in one dimension is the ratio of variances of both sets. For multiple dimensions we look for the direction vector v_0 in which this ratio is maximized, given by the following:

$$\underset{v_0}{argmax}\ \frac{v_0^T \cdot \Sigma(z_{0,1}) \cdot v_0}{v_0^T \cdot \Sigma(z_0) \cdot v_0} \tag{4}$$

where $\Sigma(z_{0,1})$ and $\Sigma(z_0)$ are the covariances of $z_{0,1}$ and z_0 respectively. This form can be recognized as a generalized Rayleigh quotient with the solution given by the largest eigenvalue and corresponding eigenvector of $\Sigma(z_0)^{-1} \cdot \Sigma(z_{0,1})$. To obtain more than one direction vector, all of which are orthogonal to each other, we solve the eigenproblem for the related symmetric matrix: $\Sigma(z_{0,1}) \cdot \Sigma(z_0)^{-1} \cdot \Sigma(z_0)^{-1} \cdot \Sigma(z_{0,1})$. This method which we call **symmetric covariance ratio projection (SCRP)**, is repeated for z_1. Preliminary tests gave slightly better separation for the first few components of SRCP for z_0 and z_1 than projecting onto the first few PCA components

of z_0 and z_1. For the third subset, we use z-scored normalized T-F bin neighborhood and apply SCRP. The number of different neighborhood windows and corresponding window sizes for all subsets are free parameters of MLRF.

2.6. Feature extraction for subsequent layers

Subsequent layers train on an equal number of incorrectly classified samples and correctly classified samples from the previous layer. In each of these two groups, there are an equal number of samples with labels 0 and 1. Features for the subsequent layers contain the three subsets for the first layer as well as two additional subsets (See Figure 1) viz. the neighborhood of the binary mask provided by the previous layer, and SCRP applied to the predicted binary mask neighborhood.

2.7. Binary mask predictions

In the testing phase, the trained MLRF classifiers are used to predict the binary mask, BM_T, an approximation of the true IBM_T of a target, X_T. We compute the estimates $\hat{X}_T$ of X_T as follows $\hat{X}_T = BM_T \circ Z_T$, where $\circ$ is the Hadamard product. The estimate $\hat{Y}_T$ for the other source Y_T is $\hat{Y}_T = (1 - BM_T) \circ Z_T$. These estimates can be transformed to audio signals by using an inverse STFT [21].

3. Experimental setup

All simulations were run using MATLAB. Source separation is evaluated on speech data from the GRID corpus database [23] and artificially generated pink noise. All audio signals are sampled at 16 kHz. The performance of all methods are evaluated on 3 groups of audio mixtures viz. Male-Female, Male-Male and Male-Pink noise (5 pairs for each set). For every speech source in any mixture we use 100 sentences from the GRID corpus database. The training data comprises 80 of these sentences which are roughly 2 minutes and 10 seconds of clean speech for each of the speakers. The testing data comprises the remaining 20 sentences which are concatenated into roughly 40 seconds of clean speech. We choose a normalization window width of 1s which is longer than typical human speech temporal patterns [24]. This normalization window width led to higher SNR gains for all methods in comparison to the results for an RMS based normalization over the entire data duration (data not shown). Artificial test mixtures are generated by linearly adding the two test audio sources at an input SNR of 0 dB. To compute the spectrograms, a Fourier window size of 1024 samples (equivalent to 64 ms) and 75% overlap between successive Hanning windows are used. Each spectrogram column thus contains 513 distinct frequency bands.

Optimal parameters for all methods were chosen by maximizing the output SNR on a single randomly selected male-female pair. For MLRF, the number of random splitting dimensions chosen for each splitting node (m_{try}), the number of trees, the number of RF layers (l) and corresponding number of training samples in each layer (n_s^l) were fixed at 64, 300, 2 and $\{n_s^1 = 15000; n_s^2 = 7000\}$ respectively. Adding more trees to the random forests or using more training samples never resulted in a decrease of classification performance. Adding more layers did not give better performance, as sampling an equal number of correctly and incorrectly classified examples exhausts the training data more quickly. Adding the second layer did not improve the performance for the Male-Pink noise case possibly due to overfitting. For the first layer we chose nine different neighborhood windows of various size combinations according to Section 2.5 and chose four neighborhood windows for the second layer. The optimal parameters for RNN were chosen by maximizing the output SNR and are in agreement with [13]. The number of layers and nodes per layer were chosen to be 3 and 1000 respectively. The optimal parameter values for NNSC were chosen as in [21] which were also obtained by maximizing the SNR.

4. Results and discussion

The source separation performance is evaluated using 4 metrics viz. Signal to Noise Ratio (SNR) [21], classification accuracy as percentage of correctly classified T-F bins between BM_T and IBM_T, PESQ [25] and STOI [26]. The SNR values tell us how close the reconstructed signal is to the target signal. The PESQ score reflect the quality of the reconstructed speech while the STOI score reflects its intelligibility. For the male-female case, the typical power overlap in a frequency band is small compared to that in the male-male case. Therefore, in the male-female case we expect a good separation based on frequency specialized classifiers while the male-male case requires better recognition of local, source specific T-F patterns for good separation. The male-pink noise case poses a different challenge, as signals typically overlap more strongly in individual T-F bins and therefore it is difficult to separate for any of the binary classification methods. The results are summarized in Table 1.

Though MLRF gives the highest mean scores for SNR (except the male-male case), both MLRF and RNN significantly outperformed NNSC methods. The classification accuracy for MLRF was found to be significantly better than other methods except for the male-pink noise case. A significantly high PESQ score for NNSC over all other binary classification methods in the male-pink noise case was observed. This possibly indicates that application of binary classification methods might be detrimental to the speech quality when applied on a noisy mixture. It has been shown that an improved quality of the reconstructed speech (higher PESQ scores), might not necessarily lead to improvements in speech intelligibility [27]. In fact, in the same study it is shown that some speech processing algorithms which achieve a significant improvement in quality might be accompanied by a decrease in intelligibility. The STOI metric has been shown to provide very high correlation (>0.9) with intelligibility scores provided by humans [26], [28], and is therefore also evaluated. MLRF performed significantly better against all other methods in all 3 mixture cases. The use of multiple metrics for evaluation allows us to conclude on different aspects of the performance of the different methods.

Different frequencies can have different roles in the perception of sound [29], [30]. Therefore, frequency-specific analysis of the results was also performed to evaluate performances of the various methods. Mean squared error computed between the target and estimated frequency bands (data not shown) indicated better performance of MLRF over all other methods especially in the higher frequency regions (>1 KHz). MLRF uses an equal number of samples for both sources in each frequency band. As a result, high recall was achieved for the target even when its power fraction (fraction of the total power of a target in a mixture for a frequency band) was low. This holds true for both sources. This is depicted in Figure 2.

Table 1. *Performance of all algorithms in regards to different metrics for all 3 mixture cases. The best performance is shown in bold. In case a method performs significantly better ($p < 0.05$) than all others in a two tailed t-test, the result is marked with a *. Here MF: Male-Female, MM: Male-Male, MN: Male-Pink Noise*

Method	Mean SNR (dB) (MF, MM, MN)	Mean classification accuracy (%) (MF, MM, MN)	Mean PESQ score (MF, MM, MN)	Mean STOI score (MF, MM, MN)
MLRF	(**10.32**, 8.01, **8.60**)	(**78.10***, **77.62***, **92.59**)	(**2.30***, 2.00, 1.57)	(**0.87***, **0.86***, **0.79***)
Deep RNN	(10.24, **8.72**, 8.52)	(60.43, 62.37, 88.17)	(1.98, 1.70, 1.57)	(0.83, 0.83, 0.76)
NNSC	(7.24, 5.11, 6.51)	Not applicable	(2.15, **2.06**, **1.98***)	(0.81, 0.79, 0.77)
NNSC (BM)	(7.51, 4.59, 7.17)	(66.5, 63.38, 86.71)	(2.03, 1.79, 1.63)	(0.80, 0.76, 0.74)

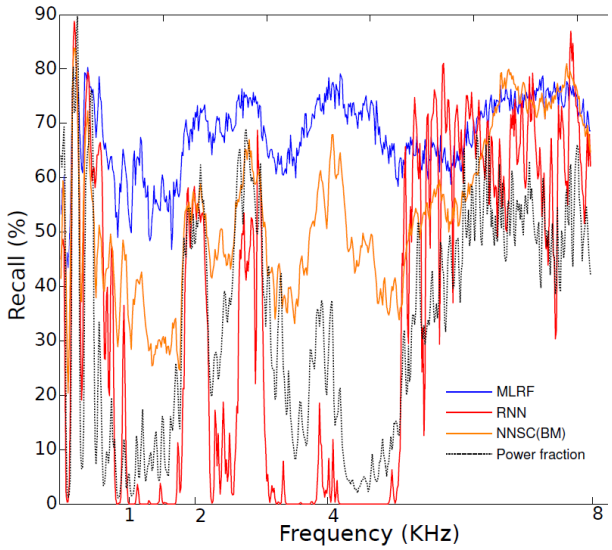

Figure 2: *Frequency-wise comparison of the recall for MLRF, RNN and NNSC (BM). Shown as a dotted black line is the frequency-specific power fraction for this target source. MLRF uses equal number of samples for both sources which helps to preserve the non-dominant source. The RNN approach in contrast follows the power fraction curve and has lower recall for lower power fraction. In some frequency bands, recall is as low as 0% leading to the assignment of the mixture to the non-target.*

Note that the figure 2. is shown only for one source. RNN in contrast follows the power fraction curve, resulting in low recall for low power fraction. A common criticism about methods based on neural networks is that they behave like a black box approach [31]. MLRF, in contrast, gives interpretable results yielding how the mixture space is partitioned at each step (node of the tree) therefore giving a deeper insight and transparency to the classification task.

Other advantages of RF-based classification is that adding more trees does not result in overfitting [22]. Also, because RF classifiers for each frequency band and for each tree are learned independently of each other, each layer of the algorithm is easily parallelizable and viable for real-time source separation.

Studies have shown that IBM in comparison to other masking methods improves speech intelligibility but not the quality of the reconstructed speech in a speech-noise mixtures [32]. In agreement with this observation, we found that for male-pink noise mixture, the NNSC approach achieved a higher PESQ score than the binary classification methods. As a possible workaround, MLRF can be extended to regression-based random forests to generate soft masks which might better address this problem.

It has been shown that humans can achieve nearly perfect speech perception when speech is demixed using IBMs [33], which is also supported by our high STOI scores. Human speech exhibits structure on both temporal [34], [24] and spectral scales [35], which helps in human auditory processing [34]. The temporal width in our MLRF approach encompasses typical patterns in speech [24] and a spectral width that covers typical human pitch and a few harmonics [35].

5. Conclusion and future work

In summary, this paper presents a novel RF-based approach for solving the monaural source separation problem. The MLRF method learns RF classifiers for individual frequency bands and approximates IBMs for each of the sources. A new sampling approach for efficiently drawing training samples is introduced. The input features exploit the properties of speech structure as they include the typical range of temporal and spectral patterns of speech. Overall, our method outperformed state-of-the-art approaches such as deep learning based on RNNs [13] and NNSC [17] on metrics covering SNR gain, classification accuracy, PESQ and STOI scores. Extensions to this work include MSS on more than two audio sources and its real-time implementation.

6. Acknowledgements

We gratefully acknowledge Dr. Florian Blättler for his useful suggestions and comments. This work was partially funded by Swiss National Science Foundation grants #200021-126844 “Early Auditory Based Recognition of Speech”, #200020-153565 “Fast Separation of Auditory Sounds” and supported by the Eth grant no 2-73246-08 and by SNF Sinergia. Cosimo Riday and Saurabh Bhargava contributed equally to the work and are the first authors of this paper.

7. References

[1] E. C. Cherry, "Some experiments on the recognition of speech, with one and with 2 ears," *The Journal of the Acoustical Society of America*, vol. 25, pp. 975–979, 1953.

[2] K. Kokkinakis and P. C. Loizou, "Using blind source separation techniques to improve speech recognition in bilateral cochlear implant patients," *The Journal of the Acoustical Society of America*, vol. 123, no. 4, pp. 2379–2390, 2008.

[3] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, vol. 26, no. 1. Wiley-Interscience, 2001.

[4] R. Vigario, J. Sarela, V. Jousmaki, M. Hamalainen, and E. Oja, "Independent component approach to the analysis of EEG and MEG recordings," *IEEE Transaction on Biomedical Engineering*, vol. 47, pp. 589–593, 2000.

[5] E. W. Healy, S. E. Yoho, Y. Wang, and D. Wang, "An algorithm to improve speech recognition in noise for hearing-impaired listeners.," *The Journal of the Acoustical Society of America*, vol. 134, no. 4, pp. 3029–38, 2013.

[6] M. S. Pedersen, "Source Separation for Hearing Aid Applications," *Ph.D. Dissertation* Informatik og Matematisk Modelling*, DTU*, 2006.

[7] G. J. Wang, D. L., and Brown, *Computational Auditory Scene Analysis Chapter 1.pdf*. Wiley and Sons, Hoboken, NJ, 2006.

[8] D. S. Brungart, P. S. Chang, B. D. Simpson, and D. Wang, "Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation.," *The Journal of the Acoustical Society of America*, vol. 120, no. 6, pp. 4007–4018, 2006.

[9] K. Han and D. Wang, "A classification based approach to speech segregation," *The Journal of the Acoustical Society of America*, vol. 132, no. 5, pp. 3475–3483, 2012.

[10] D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*. Springer US, pp. 181–197, 2005.

[11] Y. Li and D. Wang, "On the optimality of ideal binary time-frequency masks," *Speech Communications*, vol. 51, no. 3, pp. 230–239, 2009.

[12] G. J. Brown and D. Wang, "16 Separation of Speech by Computational Auditory Scene Analysis," *Speech Enhancement*, vol. 7, pp. 371–402, 2005.

[13] P. Huang, M. Kim, M. Johnson and P. Smaragdis, "Department of Electrical and Computer Engineering , University of Illinois at Urbana-Champaign , USA Department of Computer Science , University of Illinois at Urbana-Champaign , USA Adobe Research , USA," in *Deep Learning for Monaural Speech Separation*, 2014.

[14] A. Narayanan, "Investigation of speech separation as a front-end for noise robust speech recognition," ACM/*IEEE Transactions on Audio, Speech and Language Processing*, pp. 826-835, 2014.

[15] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 7092–7096, 2013.

[16] P. C. Loizou, *Speech enhancement: theory and practice*. CRC Press, Florida, 2013.

[17] M. N. Schmidt, J. Larsen, and F.-T. Hsiao, "Wind noise reduction using Non-negative Sparse Coding," *Proceedings of the IEEE Workshop on Machine Learning for Signal Processing,* pp. 431–436, 2007.

[18] D. Kitamura, H. Saruwatari, K. Shikano, K. Kondo, and Y. Takahashi, "Music signal separation by supervised nonnegative matrix factorization with basis deformation," *International Conference on DSP*, pp. 1–6, Jul. 2013.

[19] P. Sen 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.

[20] D. A. Schwartz, C. Q. Howe, and D. Purves, "The statistical structure of human speech sounds predicts musical universals.," *The Journal of Neuroscience*, vol. 23, no. 18, pp. 7160–7168, 2003.

[21] S. Bhargava, F. Blattler, S. Kollmorgen, S.C. Liu, and R. Hahnloser, "Linear Methods for Efficient and Fast Separation of Two Sources Recorded with a Single Microphone," *Neural Computation*, 2015.

[22] L. Breiman, "Random forests," *Machine Learning*, vol 45, no. 1, pp. 5–32, 2001.

[23] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition.," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[24] S. Rosen, "Temporal information in speech: acoustic, auditory and linguistic aspects.," *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, vol. 336, no. 1278, pp. 367–73, 1992.

[25] H. Yi and P. C. Loizou, "Evaluation of Objective Quality Measures for Speech Enhancement," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, pp. 229–238, 2008.

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

[27] R. L. Goldsworthy and J. E. Greenberg, "Analysis of speech-based Speech Transmission Index methods with implications for nonlinear operations.," *The Journal of the Acoustical Society of America*, vol. 116, no. 6, pp. 3679–3689, 2004.

[28] A. M. Gómez, B. Schwerin, and K. Paliwal, "Improving objective intelligibility prediction by combining correlation and coherence based methods with a measure based on the negative distortion ratio," *Speech Communications*, vol. 54, no. 3, pp. 503–515, 2012.

[29] O. Ghitza, "Processing of spoken CVCs in the auditory periphery. I. Psychophysics.," *The Journal of the Acoustical Society of America*, vol. 94, no. 5, pp. 2507–2516, 1993.

[30] O. Ghitza, "Auditory models and human performance in tasks related to speechcoding and speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 1, 1994.

[31] J. M. Benitez, J. L. Castro, and I. Requena, "Are artificial neural networks black boxes?," *IEEE Transactions on Neural Networks*, vol. 8, no. 5, pp. 1156–1164, 1997.

[32] I. Brons, R. Houben, and W. A. Dreschler, "Perceptual effects of noise reduction by time-frequency masking of noisy speech," *The Journal of the Acoustical Society of America*, vol. 132, no. 4, p. 2690, 2012.

[33] D. Wang, U. Kjems, M. S. Pedersen, J. B. Boldt, and T. Lunner, "Speech perception of noise with binary gains.," *The Journal of the Acoustical Society of America*, vol. 124, no. 4, pp. 2303–2307, 2008.

[34] A. S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*, vol. 19, no. 2. MIT Press, 1990.

[35] I. R. Titze and D. W. Martin, "Principles of voice production," *The Journal of the Acoustical Society of America*, vol. 104, no. 3, pp. 1148-1148, 1998.