Group Delay based Music Source Separation using Deep Recurrent Neural Networks

Jilt Sebastian and Hema A. Murthy
Department of Computer Science and Engineering
Indian Institute of Technology Madras, Chennai, India- 600036
Email: jiltsebastian@gmail.com, hema@cse.iitm.ac.in

***Abstract*—Deep Recurrent Neural Networks (DRNNs) have been most successfully used in solving the challenging task of separating sources from a single channel acoustic mixture. Conventionally, magnitude spectra are being used to learn the characteristics of individual sources in such monaural blind source separation (BSS) task. The phase spectra which inherently contain the timing information is often ignored.**

In this work, we explore the use of modified group delay (MOD-GD) function for learning the time-frequency masks of the sources in the monaural BSS problem. We demonstrate the use of MOD-GD through two music source separation tasks: singing voice separation on the MIR-1K data set and vocal-violin separation on the Carnatic music data set. We find that it outperforms the state-of-the-art feature in terms of Signal to Interference Ratio (SIR). Moreover, training and testing times are significantly reduced (by 50%) without compromising on the performance for the best performing DRNN configuration.

I. Introduction

Deep Neural Networks (DNNs) have gained considerable interest in recent years in acoustic modeling. As a learning approach, DNNs do not require any task-specific assumptions and prior source knowledge which may not be always true in the real world applications. The network parameters are directly learned from the data. For many of the audio applications, state-of-the-art results are obtained using deep learning [1], [2].

Monaural source separation is ill-posed and hence a challenging problem. DNN has been applied recently to BSS problems with different model architectures [3], [4] where the models learn the mapping between the mixture signal and the separated signals. Huang et al. proposed Deep Recurrent Neural Network (DRNN) for monaural Blind Source Separation (BSS) [4] in which both the sources are simultaneously modeled. Time-frequency masking is employed to make the sum of the prediction results equal to that of the original mixture. In [5], Long-Short Term Memory (LSTM) DRNNs are introduced for source separation of speech signals.

These networks are modeled to learn the time-frequency patterns for each of the sources from the raw mixture signal. Separability of these patterns in the feature domain enhances the source separation quality. At present, magnitude spectrum based features such as Mel Frequency Cepstral Coefficients (MFCC), logMel [4], [6] and the magnitude spectrum itself [7], [8], [6] are used to learn the optimum time-frequency mask. In [4], MFCC features that are commonly used for other audio applications are employed, while in [9], logMel features are used owing to the success of logMel features in Automatic Speech Recognition (ASR) [10]. However, the performance was better for the magnitude spectrum feature compared to MFCC and logMel features [6].

978-1-5090-1746-1/16/$31.00 © 2016 IEEE

For music source separation, spectrum as a feature has yielded the most promising results. When the individual pitch trajectories overlap or the formants of the different sources are closer, performance degrades and is reflected in a lower Signal to Interference Ratio (SIR). Phase spectrum based group delay function has been successfully used in Music Information Retrieval (MIR) tasks such as tonic identification [11], musical onset detection [12] and melody mono pitch extraction [13].

In this paper, we propose the phase-based Modified Group Delay (MOD-GD) feature [14], for learning the time-frequency mask in BSS as opposed to conventional magnitude spectrum based features. Features based on MOD-GD function have been used for speaker verification and it is observed in [15] that MOD-GD is the preferred feature to MFCC for a large number of speakers. Clearly, the timbre of the speaker is captured by this feature. The sources correspond to different timbres in the source separation problem. We explore the mod-gdgram feature obtained by concatenating MOD-GD function over the consecutive frames in DRNN architecture [6] and discuss the performance and the computational/architectural advantages over the spectrum feature.

The organization of this paper is as follows: Section II provides an overview of the DRNN architecture used in BSS and the modified group delay. Section III describes the proposed method using MOD-GD-gram. Section IV discusses experimental settings and results. Section V contains conclusion and the future work.

II. Related Works

A. DRNNs

Recurrent neural networks (RNN) are characterized by temporal connections between the layers of two neural networks. These are used to capture the contextual information among the sequential data. However, the hierarchical processing is

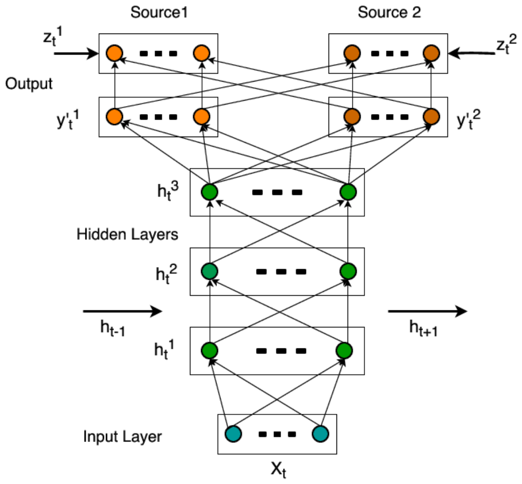

Fig. 1: DRNN architecture used for source separation (Redrawn from [6])

limited owing to the system lacking hidden layers. DRNNs provide this information at multiple time scales. Figure 1 shows a typical deep recurrent neural network architecture used in BSS [6]. l-DRNN is the one with a temporal connection at l^{th} layer. The temporal connection is present at every layer of the stacked DRNN. For an l-DRNN, the hidden activation at level l and time t is given by:

$$\begin{aligned} h_t^l &= f_h(x_t, h_{t-1}^l) && (1)\\ &= \phi_l(U_l h_{t-1}^l) + W^l \phi_{l-1}(W^{l-1}(\ldots\phi_1(W^1 x_t)))), && (2)\end{aligned}$$

The output value y_t is then obtained as,

$$\begin{aligned} y_t &= f_0(h_t^l) && (3)\\ &= W^L \phi_{L-1}(W^{L-1}(\ldots\phi_l(W^l h_t^l))), && (4)\end{aligned}$$

where x_t is the input to the network at time t, W^l is the weight matrix for the l^{th} layer, U^l is the weight matrix for the recurrent connection at the l^{th} layer and $\phi_l(\cdot)$ is the nonlinear activation function. Huang et al. [6] empirically found that the rectified linear unit $f(x) = max(0,x)$ performs better compared to using a sigmoid or $tanh$ activation function.

Feature vector x_t is given as the input to the network to obtain the source estimates, $y_t'^1$ and $y_t'^2$. The soft time-frequency mask [16] is applied to the magnitude spectrum of the mixture signal to obtain the separated spectra (z_t^1 and z_t^2). This masking function is added as an additional deterministic layer and the network is jointly optimized with the masking function. The network parameters are optimized by minimizing the Mean Squared Error (MSE) objective function and Kullback-Leibler divergence (KL) criteria. This discriminative objective function not only increases the similarity between the prediction and target but also decreases the similarity between the prediction and the targets of other sources.

The objective function is given by:

$$||\hat{y}_{1t} - y_{1t}||_2^2 - \gamma||\hat{y}_{1t} - y_{2t}||_2^2 + ||\hat{y}_{2t} - y_{2t}||_2^2 - \gamma||\hat{y}_{2t} - y_{1t}||_2^2 \quad (5)$$

and the divergence criteria used is:

$$D(y_{1t}||\hat{y}_{1t}) - \gamma D(y_{1t}||\hat{y}_{2t}) + D(y_{2t}||\hat{y}_{2t}) - \gamma D(y_{2t}||\hat{y}_{1t}), \quad (6)$$

where $D(A||B)$ is the KL divergence between A and B. The γ parameter is chosen based on development data performance.

B. *Modified group delay*

Audio attention in humans is related to timing. It is conjectured that source separation may be better modeled using features wherein the timing information is preserved. i.e, the *phase* spectrum. The group delay, defined as the negative derivative of phase with respect to frequency, is used as an alternative to the phase spectrum. The ratio of the peak amplitude to the amplitude at 3 dB bandwidth (as defined by the magnitude spectrum) is always higher for the group delay function compared to that of the magnitude spectrum [17]. This high-resolution capability of the group delay function resolves formants and pitch better. Figure 2 illustrates this property for a sum of two sinusoids. Observe that the sinusoidal peaks are visible in the group delay spectrum even at low Signal to Noise Ratios (SNRs). Due to windowing in the short-time analysis, zeroes are introduced close to the unit circle in the Z-domain and they appear as peaks in the group delay function. The modified group delay function was proposed to reduce this effect.

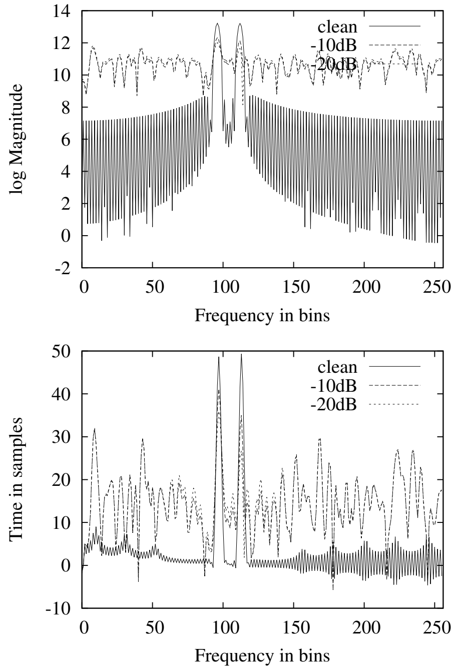

Fig. 2: Magnitude spectrum (top) and the group delay function (bottom) of sum of two sinusoids at different noise levels

The modified group delay function of a discrete time signal $x[n]$ with its Fourier transform $X(\omega)$ can be computed [18] as:

$$\tau_m(\omega) = \left(\frac{\tau(\omega)}{|\tau(\omega)|}\right)(|\tau(\omega)|)^{\alpha_i} \quad (7)$$

where,

$$\tau(\omega) = \Big(\frac{X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega)}{|S(\omega)|^2}\Big), \tag{8}$$

$Y(\omega)$ is the Fourier transform of $n.x[n]$, $|S(\omega)|^2$ is the smoothed version of $|X(\omega)|^2$. The first multiplicative term in equation 7 is the sign of the modified group delay (+1/-1) and α_i is a parameter that is used for controlling the dynamic range of this feature, with $i = 1$ and 2 for positive and negative scales respectively. These positive and negative scales determine the emphasis given to the positive and negative values of the MOD-GD function. The information contained in the phase spectrum is utilized for applications such as pitch estimation, formant estimation, and speaker recognition by using the modified group delay function or features derived from it [19]. However, it has not been employed as a feature in the source separation task so far. Modified group delay is used for obtaining the pitch estimates in [13] and is extended in [20] for multi-pitch estimation since the modgdgram shows prominent peaks at both of the pitch periods and its harmonics.

III. BSS WITH MODGDGRAM

The architecture of DRNN shown in Figure1 is used with the MOD-GD feature for music source separation. The input feature to the DRNN network is the modified group delay-gram (modgdgram) which is obtained by concatenating MOD-GD function of the successive frames. The time-frequency mask learned from them are used to filter the mixture magnitude spectrum to obtain the individual source spectra. The MOD-GD is computed from the signal and its time weighted version, as given in equation 7 and 8. In this work, the moving average smoothing function is used in place of cepstral smoothing function [19] as the former is more robust to zeros in the frequency domain. As regions around the formants are important for timbre, the powers for the positive peaks (α_1) are set different from that of the negative peaks (α_2).

Figure 3 compares the spectrogram and the modgdgram of the sources and their linear mixtures used in singing voice separation for a music segment from the MIR-1K dataset. The time-frames are squeezed to make the pitch trajectories visible. FFT size is chosen to be 512 and the lower 100 bins are used for plotting since it has most of the melodic information. It should be noted that the mixture modgdgram preserves the harmonics of the sources better than the mixture spectrogram. Observe from the figure that the dynamic range is higher for the modgdgram compared to that of the spectrum, in that pitch trajectory stands out with respect to the background. The MOD-GD feature has a comparable computational complexity to that of the spectrum for the same input dimension.

IV. EXPERIMENTS AND RESULTS

We evaluate the source separation performance using the MOD-GD feature on two music source separation tasks: singing voice separation and vocal-violin separation. 3 layer DRNN architecture with discriminative objective function (Equation 5) is used in the experiments. We set the maximum epoch to 400 in each configuration.

A. Evaluation Metrics

The source separation quality is measured using three quantitative measures based on BSS-EVAL 3.0 metrics [21]: Source to Artifacts Ratio (SAR), Source to Interference Ratio (SIR) and Source to Distortion Ratio (SDR). The amount of suppression achieved for the interfering source is represented in SIR which is an indicator of the timbre differences between two sources. Normalized SDR (NSDR) is defined by [6] as:

$$NSDR(\widehat{v}, v, x) = SDR(\widehat{v}, v) - SDR(x, v), \tag{9}$$

where x is the mixture, $\widehat{v}$ and v are the estimated source and the actual clean source respectively. Improvement of the SDR between the mixture and the separated source is reflected in NSDR. The Test clips are weighted by their length and their weighted means are used to represent the overall performance via Global SAR (GSAR), Global SIR (GSIR) and Global NSDR (GNSDR).

B. Datasets used

For the singing voice separation task, the MIR-1K dataset [22] is used to evaluate the performance of the MOD-GD feature. It consists of thousand song clips at 16 kHz sampling rate with durations ranging from 4 to 13 seconds. Each clip contains the singing voice and the background music in different channels. These clips were extracted from 110 Chinese karaoke songs performed by male and female amateurs. Training set consists of 171 clips sung by one male and one female (”abjones” and ”amy”). The development set contains 4 clips sung by the same singers, following the same framework as in [6]. The test set consists of the remaining 825 clips from 17 amateurs. Channels are mixed at 0 dB SNR and our aim is to separate the singing voice from the background music.

Since there was no dataset specifically for Carnatic music source separation, we have created a datset ourselves for vocal-violin separation task. From a concert of 2 hours and 3 minutes duration, 77 musical clips are extracted with the duration ranging from 2 to 23 seconds. The recorded data is a two channel signal with the vocal in one channel and the lead instrument (violin) in the other. These are mixed at equal energy levels to obtain a single channel mixture signal. The training data consists of randomly selected 54 clips, the development set contains 3 clips and the test set consists of remaining 20 clips.

C. Singing voice separation in MIR-1K dataset

Experiments are performed with both the modgdgram and magnitude spectrogram features. The spectral representation is extracted using 1024 point short time Fourier transform (STFT) with an overlap of 50%. Following [6], we have used a 32ms window with 16ms frame shift for calculating the features. Since the context features can further improve the performance, we have used a contextual window of 3 frames. In the modified group delay computation, smoothing parameter is set to 5 and the group delay scales (α_i) are set to 1.2 and 0.45, as obtained from the multi-pitch task [20].

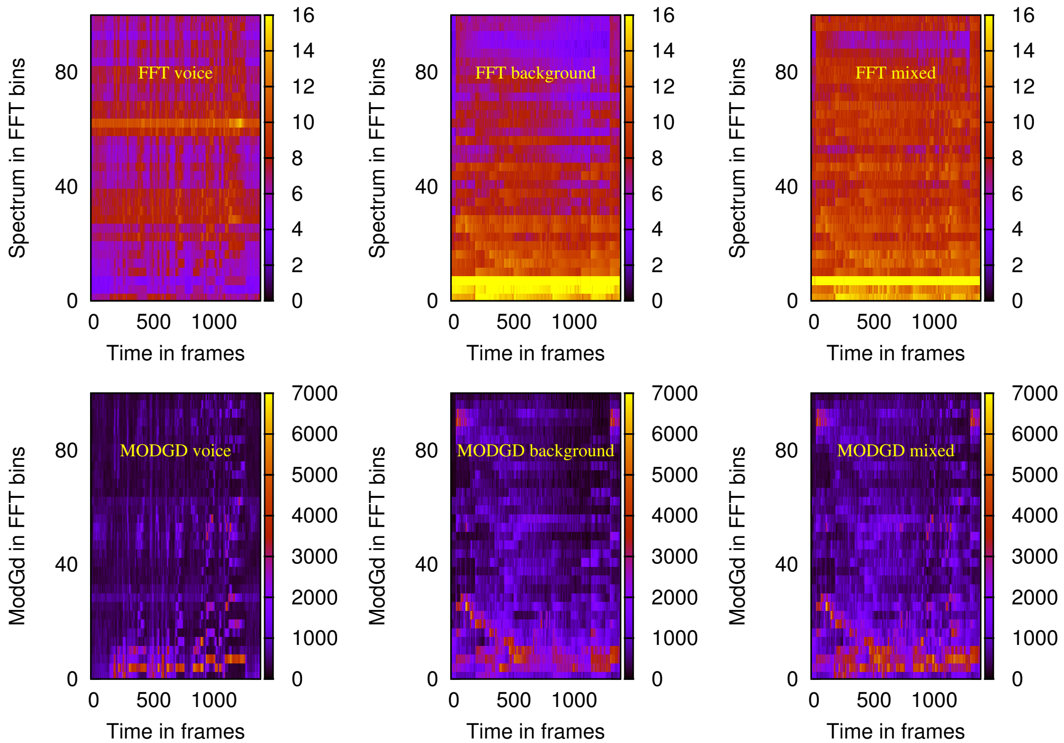

Fig. 3: Feature representations of the clip Ani_1_01.wav from MIR-1K dataset. The first row represents the spectrogram and second row represents log-modgdgram. Third column is the linear mixture of the first (singing voice) and second (background music) columns.

The performance of the MOD-GD feature is compared with that of the magnitude spectrum feature on several aspects. In terms of complexity (Table I), it is observed that the architecture with just 500 hidden nodes per layer performs similar to that of the architecture with 1000 nodes per layer with the spectrum feature. Hence, a network with 1500 fewer hidden nodes is sufficient to achieve the same performance, i.e, training and testing times are halved.

TABLE I: Performance measures with 2-DRNN

Feature	**Hidden units per layer**	**GNSDR**	**GSIR**	**GSAR**
ModGD	500	**7.15**	**13.46**	**9.11**
Spectrum	500	5.74	12.15	7.62
ModGD	1000	**7.50**	**13.73**	9.45
Spectrum	1000	7.45	13.08	9.68

We also compare the best results (2-DRNN) obtained using the spectrum feature [6] with our approach in Table I. For the same setting, modgdgram feature gives similar results for SAR and SDR and shows a relative improvement of 4.9%dB for SIR over magnitude spectrum. This is because the mask is learned from the group delay domain, where the resolution is higher than the spectrum. Note that there is not much improvement from 500 to 1000 hidden units per layer, which suggests intelligent separation is possible with a simpler network with modgdgram feature.

TABLE II: Results with DRNN architectures

Architecture	**Feature**	**GNSDR**	**GSIR**	**GSAR**
1-DRNN	Spectrum	7.21	12.76	**9.56**
	ModGD	**7.26**	**12.93**	9.42
2-DRNN	Spectrum	7.45	13.08	**9.68**
	ModGD	**7.50**	**13.73**	9.45
3-DRNN	Spectrum	**7.09**	11.69	**10.00**
	ModGD	6.92	**12.27**	9.26
stacked DRNN	Spectrum	7.15	12.79	**9.39**
	ModGD	**7.31**	**13.45**	9.30

Table II shows the performance of the feature on several RNN configurations compared to the spectrum. Better SIR ratio is achieved for *all* the configurations with similar values for other measures. Thus, modgdgram improves the quality of separation irrespective of the model configurations.

D. Vocal-Violin separation in Carnatic music dataset

Carnatic music is a particular classical form performed in the southern region of India. In a concert, the vocal and all the accompanying instruments are tuned to the same base frequency called *tonic* frequency. This can lead to overlapping of the pitch frequencies corresponding to vocal and other instruments. Hence, Carnatic music source separation is not possible with simple dictionary learning methods. This is the first attempt at source separation for a live Carnatic music concert with no constraint on the data.

We compare the results obtained with modgdgram and spectrogram features on an architecture with 1000 hidden units per layer. The architecture of DRNN with a temporal connection at 1^{st} hidden layer (1-DRNN) is used to obtain the results. Other experimental settings are made similar to that of singing voice separation task. From Table III, it is observed that the performance of both the features are almost equal, with modgdgram feature giving slightly better GSIR. This is also reflected in the GNSDR.

TABLE III: 1-DRNN performance in the Carnatic music dataset.

Feature	GNSDR	GSIR	GSAR
ModGD	**9.42**	**13.72**	11.76
Spectrum	9.38	13.55	**11.80**

From the experiments it can be inferred that the modgdgram can replace the spectrogram feature for the music source separation task in the state-of-the-art DRNN architecture because of two major reasons: First, it gives better GSIR values and second, the modgdgram based DRNN is less complex, resulting in a reduction of the computation time by 50% in the best configuration of the architecture. We also conjecture that the higher resolution property helps in learning the average time-frequency trajectories with a simpler network.

V. Conclusion and Future work

In this paper, we propose the use of phase based modgdgram feature with the deep recurrent learning models for music source separation from monaural recordings. The proposed modgdgram feature achieves improved results with respect to GSIR on all the architectures while maintaining the state-of-the-art performance with respect to GSARs and GNSDRs and also requires a less complex DRNN configuration for similar performance. Our future work will include applying the proposed feature for speech separation and speech denoising tasks. Since the modgdgram offers higher resolution, the need for discriminative training will also be analyzed.

VI. Acknowledgment

This research is partly funded by the European Research Council under the European Unions Seventh Framework Program, as part of the CompMusic project (ERC grant agreement 267583). Authors would like to thank Dr.Padmasundari for suggesting the changes in the draft.

References

[1] G. Hinton, Li Deng, Dong Yu, G. E Dahl, Abdel-rahman Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, Tara N Sainath, et al., “Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups,” *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 82–97, 2012.

[2] Xavier Glorot, Antoine Bordes, and Yoshua Bengio, “Deep sparse rectifier neural networks,” in *International Conference on Artificial Intelligence and Statistics*, 2011, pp. 315–323.

[3] Nicolas Boulanger-Lewandowski, Gautham J Mysore, and Matthias Hoffman, “Exploiting long-term temporal dependencies in nmf using recurrent neural networks with application to source separation,” in *Acoustics, Speech and Signal Processing (ICASSP), 2014*. IEEE, 2014, pp. 6969–6973.

[4] Po-Sen Huang, Minje Kim, Mark Hasegawa-Johnson, and Paris Smaragdis, “Deep learning for monaural speech separation,” in *Acoustics, Speech and Signal Processing (ICASSP), 2014*. IEEE, 2014, pp. 1562–1566.

[5] Felix Weninger, John R Hershey, Jonathan Le Roux, and Bjorn Schuller, “Discriminatively trained recurrent neural networks for single-channel speech separation,” in *Signal and Information Processing (GlobalSIP), 2014 IEEE Global Conference on*. IEEE, 2014, pp. 577–581.

[6] Po-Sen Huang, Minje Kim, Mark Hasegawa-Johnson, and Paris Smaragdis, “Singing-voice separation from monaural recordings using deep recurrent neural networks,” *International Society for Music Information Retrieval (ISMIR)*, 2014.

[7] Gautham J Mysore, Paris Smaragdis, and Bhiksha Raj, “Non-negative hidden markov modeling of audio with application to source separation,” in *Latent Variable Analysis and Signal Separation*, pp. 140–148. Springer, 2010.

[8] Andrew JR Simpson, “Probabilistic binary-mask cocktail-party source separation in a convolutional deep neural network,” *arXiv preprint arXiv:1503.06962*, 2015.

[9] Po-Sen Huang, Minje Kim, Mark Hasegawa-Johnson, and Paris Smaragdis, “Joint optimization of masks and deep recurrent neural networks for monaural source separation,” *arXiv preprint arXiv:1502.04149*, 2015.

[10] Jinyu Li, Dong Yu, Jui-Ting Huang, and Yu Gong, “Improving wideband speech recognition using mixed-bandwidth training data in cd-dnn-hmm,” in *Spoken Language Technology Workshop (SLT), 2012 IEEE*. IEEE, 2012, pp. 131–136.

[11] Ashwin Bellur and Hema A Murthy, “A novel application of group delay functions for tonic estimation in carnatic music,” in *eusipco*, September 2013, pp. Th–L1.4.

[12] Manoj Kumar, Jilt Sebastian, and Hema A Murthy, “Musical onset detection on carnatic percussion instruments,” in *Communications (NCC), 2015 Twenty First National Conference on*. IEEE, 2015, pp. 1–6.

[13] R. Rajan and H.A. Murthy, “Group delay based melody monopitch extraction from music,” in *Acoustics, Speech and Signal Processing (ICASSP), 2013*, May 2013, pp. 186–190.

[14] Hema A Murthy B Yegnanarayana and V R Ramachandran, “Processing of noisy speech using modified group delay functions,” *ICASSP*, pp. pp.945–948, May 1991.

[15] T Asha, MS Saranya, DS Karthik Pandia, S. Madikeri, and Hema A Murthy, “Feature switching in the i-vector framework for speaker verification,” in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[16] Po-Sen Huang, Scott Deeann Chen, Paris Smaragdis, and Mark Hasegawa-Johnson, “Singing-voice separation from monaural recordings using robust principal component analysis,” in *Acoustics, Speech and Signal Processing (ICASSP), 2012*. IEEE, 2012, pp. 57–60.

[17] Jilt Sebastian, Manoj Kumar, and Hema A Murthy, “An analysis of the high resolution property of group delay functions with application to speech and music signals,” *Submitted to Signal Processing*, 2015.

[18] Rajesh M Hegde, Hema Murthy, Venkata Ramana Rao Gadde, et al., “Significance of the modified group delay feature in speech recognition,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, 2007.

[19] Hema A Murthy and B Yegnanarayana, “Group delay functions and its application to speech processing,” *Sadhana*, vol. 36, no. 5, pp. 745–782, November 2011.

[20] Rajeev Rajan and Hema A. Murthy, “Modified group delay based multipitch estimation in co-channel speech by adaptive filtering,” *Submitted to Signal Processing*, 2015.

[21] Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte, “Performance measurement in blind audio source separation,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 4, pp. 1462–1469, 2006.

[22] Chao-Ling Hsu and Jyh-Shing Roger Jang, “On the improvement of singing voice separation for monaural recordings using the mir-1k dataset,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 2, pp. 310–319, 2010.