INTERSPEECH 2009 BRIGHTON

A Semi-blind Source Separation Method with A Less Amount of Computation Suitable for Tiny DSP Modules

Kazunobu KONDO, Makoto YAMADA, Hideki KENMOCHI

Center for Advanced Sound Technologies, Yamaha Corp., Shizuoka, Japan

tashin@beat.yamaha.co.jp, makoto_yamada@gmx.yamaha.com, kenmochi@beat.yamaha.co.jp

Abstract

In this paper, we propose a method of implementing FDICA on *tiny* DSP modules. Firstly, we show a semi-blind separation matrix initialization step that consists of an estimation method using *covariance fitting* for a *known* source and an *unknown* source. It contributes to the faster convergence and less amount of computation. Secondly, a learning band selection step is shown that consists of the determinant of the covariance matrix as a criteria for selection; This achieves a significant reduction of an amount of computation with practical separation performance. Finally, the effectiveness of the proposed method is evaluated via the source separation simulations in anechoic and reverberant rooms, and also a procedure and a resource presumption for the integrated method which we call *tinyICA* are shown.

Index Terms: semi-blind source separation, independent component analysis, less amount of computation

10.21437/Interspeech.2009-416

1. Introduction

Today, mobile devices are so common and seen everywhere. Therefore, almost everyone expects that sound quality for telephone calls will be improved especially in noisy environments. For speech enhancement field, the source separation technique known as frequency domain independent component analysis (FDICA) [1] has received much attention from industries [2].

In [2], a real-time mobile source separation device based on FDICA is proposed. The device consists of a floating-point DSP and external memory, and it achieves reasonable separation performance. The system needs 32bit floating-point accuracy, high power consumption (300MHz clock) and large memory (about 2MByte). However, it is not realistic in practice for mobile devices, because many industries prefer to use much less-power and smaller-memory DSP (i.e., *tiny* DSP) modules typically 16bit fixed-point and under 100MHz clock with an internal 256kByte memory.

To realize a real-time operation of FDICA on such *tiny* DSP modules, we need to solve three problems: 1) optimal convergence, 2) large memory consumption and 3) computational intensiveness.

1) **Optimal convergence:** FDICA is an iterative *non-convex* approach, and thus there exist many local solutions. This fact implies that if the initial separation matrix is wrongly set, the obtained separation matrix is not guaranteed to be optimal. Concurrently, FDICA does not guarantee the order of separated source signals, and thus we cannot assign a separated target signal to a desired output channel without any prior knowledge of source signals. In other words, performance of FDICA strongly depends on the initial separation matrix.

2) **Large memory consumption:** Conventional FDICA estimates the separation matrices by storing the observed signals for each frequency bin, and thus it needs large memory.

3) **Computational intensiveness:** The non-optimal convergence mentioned in 1) increases the number of iterations, and thus it leasds to the large amount of computation. Moreover, the large memory consumption leads to an intrinsic large amount of computation which includes separation processes for every iteration.

Of course this setup is not preferable for mobile implementation, since *tiny* DSP modules have considerablely low clock and small-sized memory.

In this paper, we first propose a separation matrix initialization based on a second-order statistics to reduce the amount of iterative calculation. It alleviates the convergence and source assignment problems. We here assume a two-source separation problem, where a *known* point source, e.g., a speech signal, is placed in front of microphones, while there is no geometrical information about the other *unknown* source signal.

At the initialization step, we first estimate the direction of arrival (DOA) of the *unknown* source for each frequency and then use the DOAs for the separation matrix initialization. Here, we estimate the unknown source DOA from the covariance matrix which is estimated by using *covariance fitting* [3].

For memory consumption reduction, we propose a method of selecting the base bands for learning, based on a second-order statistics. The sound source, e.g., a speech signal, consists of a set of some predominant frequency bands, and thus it is natural to assume that limited frequency bands are sufficient for learning FDICA. Thus, we can expect that selecting the useful frequency bands for learning FDICA would contribute to reduction of memory consumption without degrading the separation performance. We use the determinant of the covariance matrix as a criteria.

Finally, the effectiveness of the proposed method is evaluated via the source separation simulations in anechoic and reverberant rooms. Because we intend to implement FDICA on *tiny* DSP modules, hereinafter, we call the system based on the proposed method *tinyICA*.

The rest of this paper is organized as follows. Section 2 explains FDICA. Section 3 introduces techniques to achieve the less amount of computation. Experimental results are reported in Section 4; Section 5 concludes with a summary of our contributions and possible future works.

2. Frequency domain ICA

In this section, we formulate signals and transmission systems, and then we explain FDICA algorithm. In this paper, we assume the number of signals and microphones are 2.

The observed signals in the time domain are transformed into the frequency domain by STFT (Short Time Fourier Transform). The convolutive model describing a propagation and a mixing of two sources in a natural environment is expressed in the frequency domain:

$$\mathbf{x}(f,\tau) = \mathbf{A}(f)\mathbf{s}(f,\tau), \tag{1}$$

where $\mathbf{x}(f,\tau) = [x_1(f,\tau), x_2(f,\tau)]^T$ is an observed signal vector at the microphones, $\mathbf{s}(f,\tau) = [s_1(f,\tau), s_2(f,\tau)]^T$ is a source signal vector, f is a frequency index, τ is a frame index and $(\cdot)^T$ denotes transpose. $\mathbf{A}(f)$ is a mixing matrix which consists of frequency responses of impulse responses, where one of elements $a_{mn}(f)$ is a complex number:

$$\mathbf{A}(f) = \begin{bmatrix} a_{11}(f) & a_{12}(f) \\ a_{21}(f) & a_{22}(f) \end{bmatrix}. \tag{2}$$

The technique known as FDICA is commonly-used for the convolutive mixture. After STFT, the separation matrix $\mathbf{W}(f)$ is obtained using infomax and natural gradient method that nonlinear function is the sign function (an approximation for the $\tanh$ function):

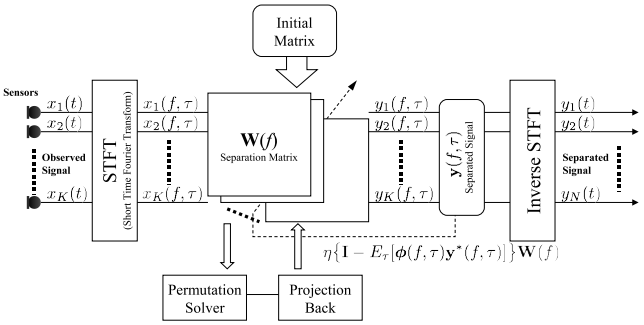

Figure 1: *Block diagram of conventional FDICA.*

$$\mathbf{y}(f,\tau) = \mathbf{W}(f)\mathbf{x}(f,\tau), \tag{3}$$

where $\mathbf{y}(f,\tau) = [y_1(f,\tau), y_2(f,\tau)]^T$ is a separated signal vector. In our experiment described in the section 4, we adopt a following learning equation:

$$\begin{aligned}\mathbf{W}_{l+1}(f) = \mathbf{W}_l(f) + \\ \eta\big\{\mathbf{I} - E_\tau[\boldsymbol{\phi}(f,\tau)\mathbf{y}^*(f,\tau)]\big\}\mathbf{W}_l(f),\end{aligned} \tag{4}$$

where l is an iteration number, $E_\tau[\cdot]$ is the expectation operator at the time frame τ and $(\cdot)^*$ is the Hermitian operator.

$$\boldsymbol{\phi}(f,\tau) \equiv [\phi_1(f,\tau), \phi_2(f,\tau)]^T, \tag{5}$$

$$\phi_k(f,\tau) \equiv \mathrm{sgn}(\Re\{y_k(f,\tau)\}) + j\ \mathrm{sgn}(\Im\{y_k(f,\tau)\}), \tag{6}$$

denotes the nonlinear function and its vector, where $\Re\{\cdot\}$ is the real part and $\Im\{\cdot\}$ is the imaginary part.

It is important to obtain the proper separation matrix that two ambiguities must be solved. The scaling ambiguity can be solved by applying the projection back method [4], and the permutation problem can be solved by using the DOA information [5].

3. Reducing the amount of computation with the second-order statistics

Repeatedly, for the mobile devices such as *tiny* DSP modules, the low amount of computation is one of the most important key issues. Hence, we use the second-order statistics in order to save the amount of computation.

The covariance matrix is the common second-order statistics and easy to obtain because of the less amount of computation than the higher-order statistics.

In addition, under the mixing model in (1), we assume two conditions: 1) the direction of one of the source signals is already known, and 2) a number of the sources is same as a number of the sensors. Hereinafter, we show our proposed method under these assumptions and using the covariance matrix.

First, the matrix initialization step based on an adaptive beamforming [3] is introduced to avoid an ill convergence to local minima, and it achieves a less number of iterations concurrently. Second, some bands are selected by the determinant of the covariance matrix, and then learning of higher-order ICA is performed. For non-selected bands, alternative separation matrices are used, and also they consist of coefficients of beamformer which directions are estimated from the separation matrices with higher-order ICA. Moreover, we show an integration procedure between these methods —we call an integrated method *tinyICA*—, and then a presumption of the amount of computation is shown.

3.1. Covariance matrix

The covariance matrix is calculated from the observed signal vector:

$$\mathbf{R}_{xx}(f) = E_\tau[\mathbf{x}(f,\tau)\mathbf{x}^*(f,\tau)], \tag{7}$$

where $\mathbf{R}_{xx}(f)$ is the covariance matrix for one frequency bin.

3.2. Matrix initialization with the covariance fitting

In this section, we explain the matrix initialization step for the two-source separation, and then we assume that one source comes from the *known* direction and the other source comes from the *unknown* direction. We call the former source the *known source* and the latter source the *unknown source*. Here, the matrix initialization step using the covariance fitting [3] is shown as follows:

For the two-source separation in anechoic, the observed signal vector is rewritten as follows:

$$\mathbf{x}(f,\tau) = s_1(f,\tau)\mathbf{v}(f,\theta_1) + s_2(f,\tau)\mathbf{v}(f,\theta_2), \tag{8}$$

where θ_1 and θ_2 are direction of the known source and unknown source. $\mathbf{v}(f,\theta_i) = [1, e^{j\rho(f,\theta_i)}]$ $(\rho(f,\theta_i) = 2\pi f d \sin(\theta_i)/V_c)$ is a representation of delays from direction θ_i in the frequency domain, d is an interval of microphones, and V_c is the sonic speed.

When the sources are uncorrelated each other, the covariance matrix for (8) is $\mathbf{R}_{xx}(f) = \mathbf{R}_{x_1x_1}(f) + \mathbf{R}_{x_2x_2}(f)$ where $\mathbf{R}_{x_1x_1}(f)$ and $\mathbf{R}_{x_2x_2}(f)$ are the covariance matrices of the known source and unknown source. Each covariance matrix is written by a source power σ_i^2 and a delay $\mathbf{v}(f,\theta_i)$, and it is expressed $\mathbf{R}_{x_ix_i}(f) = \sigma_i^2(f)\mathbf{v}(f,\theta_i)\mathbf{v}^*(f,\theta_i)$.

From the relation between $\mathbf{R}_{x_1x_1}(f)$ and $\mathbf{R}_{x_2x_2}(f)$, the covariance matrix of the unknown source is given as follows:

$$\begin{aligned}\mathbf{R}_{x_2x_2}(f) &= \mathbf{R}_{xx}(f) - \mathbf{R}_{x_1x_1}(f) \\ &= \sigma_2^2(f)\mathbf{v}(f,\theta_2)\mathbf{v}^*(f,\theta_2).\end{aligned} \tag{9}$$

Determine an array coefficient vector $\mathbf{w}(f)$ that is the solution to the eigen value problem (10), and thus $\mathbf{w}(f)$ should be equal to $\mathbf{v}(f,\theta_2)$:

$$\max\ \mathbf{w}^*(f)\mathbf{R}_{x_2x_2}(f)\mathbf{w}(f) \quad \text{s.t.}\ \mathbf{w}^*(f)\mathbf{w}(f) = 1. \tag{10}$$

Therefore, the direction of the unknown source is derived from the DOA estimation as follows:

$$\begin{aligned}\hat{\theta}_2(f) &= \arg\max_\theta |\mathbf{w}^*(f)\mathbf{d}_f(\theta)|, \\ \hat{\theta}_2 &= \frac{1}{f_s - f_e}\sum_{i=f_s}^{f_e}\hat{\theta}_2(f),\end{aligned} \tag{11}$$

where $\mathbf{d}_f(\theta_i) = [1, e^{j\rho(f,\theta_i)}]$ is a steering vector, and f_s and f_e are arbitrary frequencies.

Actually, $\mathbf{R}_{x_2x_2}(f)$ is unknown and must be estimated. So we use the covariance fitting [3]:

$$\begin{aligned}\max\quad & \sigma_1^2(f) \\ \text{s.t.}\quad & \mathbf{R}_{xx}(f) - \sigma_1^2(f)\mathbf{v}(f,\theta_1)\mathbf{v}^*(f,\theta_1) \succeq 0,\end{aligned} \tag{12}$$

where $\mathbf{M} \succeq 0$ means $\mathbf{M}$ is positive semidefinite. $\hat{\sigma}_1^2(f)$ is estimated as:

$$\hat{\sigma}_1^2(f) = \frac{1}{\mathbf{v}(f,\theta_1)\mathbf{R}_{xx}^{-1}(f)\mathbf{v}^*(f,\theta_1)}. \tag{13}$$

Then the estimated covariance matrix of the unknown source, $\hat{\mathbf{R}}_{x_2x_2}(f)$, is derived

$$\hat{\mathbf{R}}_{x_2x_2}(f) = \mathbf{R}_{xx}(f) - \hat{\sigma}_1^2(f)\mathbf{v}(f,\theta_1)\mathbf{v}^*(f,\theta_1). \tag{14}$$

Finally, the direction of the known source θ_1 and the estimated direction of the unknown source $\hat{\theta}_2$ are applied to the null beamformer as the initial separation matrix.

$$\mathbf{W}_0(f) = \begin{bmatrix}1 & -e^{j\rho(f,\hat{\theta}_2)} \\ 1 & -e^{j\rho(f,\theta_1)}\end{bmatrix}. \tag{15}$$

3.3. Learning band selection

In this section, we show that the determinant of the covariance matrix indicates source powers and a scale factor of propagations and mixing, and also it is shown as an appropriate scalar value to select useful bands for learning FDICA.

The covariance matrix is transformed by (1), and hereinafter the frequency index f and the frame index τ are omitted to simplify equations:

$$\mathbf{R}_{xx} = \mathbf{A}E_\tau[\mathbf{s}\mathbf{s}^*]\mathbf{A}^* \equiv \mathbf{A}\mathbf{R}_{ss}\mathbf{A}^*, \tag{16}$$

where $\mathbf{R}_{ss}$ is the covariance matrix of the source signals. It is assumed that $\mathbf{R}_{ss}$ can decompose the eigen values $\mathbf{\Lambda}$ and the eigen vectors $\mathbf{Q}$. Moreover, it is expressed $\mathbf{R}_{ss} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^*$ where $\mathbf{Q}$ is the orthonormal matrix.

The determinant of $\mathbf{Q}$ satisfies $\det(\mathbf{Q}\mathbf{Q}^*) = 1$, where the determinant of a matrix $\mathbf{M}$ is denoted $\det\mathbf{M}$. The diagonal matrix $\mathbf{\Lambda}$ is equal to $\mathrm{diag}\{\lambda_i\}$ where λ_i is the i-th eigen value, and thus $\det\mathbf{R}_{ss} = \det\mathbf{\Lambda}$ is derived. Moreover, the determinant of the covariance matrix is expressed as follows:

$$\begin{aligned}\det\mathbf{R}_{xx} &= \det\mathbf{A}\,\det\mathbf{\Lambda}\,\det\mathbf{A}^* \\ &= |\det\mathbf{A}|^2 \prod_i \lambda_i \quad (\because (\det\mathbf{A})^* = \det\mathbf{A}^*).\end{aligned} \tag{17}$$

The equation (17) implies a relation between sources and transmissions. The sense of $\det\mathbf{A}$ is a scaling factor of the linear projection of $\mathbf{A}$ in basic linear algebra. From (17), the covariance matrix of the observed signals shows source powers that their amounts are changed with propagation and mixing. If there is the only one source in one of the frequency bands, the rank of $\mathbf{R}_{xx}$ is not full, and $\det\mathbf{R}_{xx}$ becomes zero. Therefore, scale of $\det\mathbf{R}_{xx}$ is appropriate criteria to select the frequency bands.

These aspects imply that the determinant of the covariance matrix has information of source signals and systems of propagation and mixing.

Finally, the separation matrix for non-learning bands should be fixed. It was shown that the separation matrix of ICA is equivalent to the acoustical beamformer [6]. The DOA $\tilde{\theta}_i$ is estimated from the separation matrix similar to (11), and sets the null beamformer coefficients into the non-learning bands:

$$\mathbf{W}(f) = \begin{cases} \mathbf{W}_{ICA}(f) & (learned) \\ \mathbf{W}_{NBF}(f) = \begin{bmatrix} 1 & -e^{j\rho(f,\tilde{\theta}_2)} \\ 1 & -e^{j\rho(f,\tilde{\theta}_1)} \end{bmatrix} & (not\ learned) \end{cases}. \tag{18}$$

3.4. Integration procedure

In section 3.2 and 3.3, two new methods are introduced for saving the amount of computation of FDICA. They are integrated and described as the following process, and we call this method *tinyICA*.

1. Calculate $\mathbf{R}_{xx}$ for each frequency bin.
2. Calculate $\det\mathbf{R}_{xx}$.
3. Covariance Fitting:
   (a) Estimate the known source power $\hat{\sigma}_1^2(f)$.
   (b) Estimate the covariance matrix $\hat{\mathbf{R}}_{x_2x_2}$.
   (c) Estimate the direction $\hat{\theta}_2(f)$ and $\hat{\theta}_2$.
4. Learning Band Selection:
   (a) Select P biggest $\det\mathbf{R}_{xx}$ bins.
   (b) Learn the separation matrix for the P bins.
   (c) Estimate DOAs $\{\tilde{\theta},\tilde{\theta}_2\}$ from the separation matrices.
   (d) Set NBF with $\{\tilde{\theta},\tilde{\theta}_2\}$ for the non-selected bands.

The block diagram of the proposed method is shown in Fig-2.

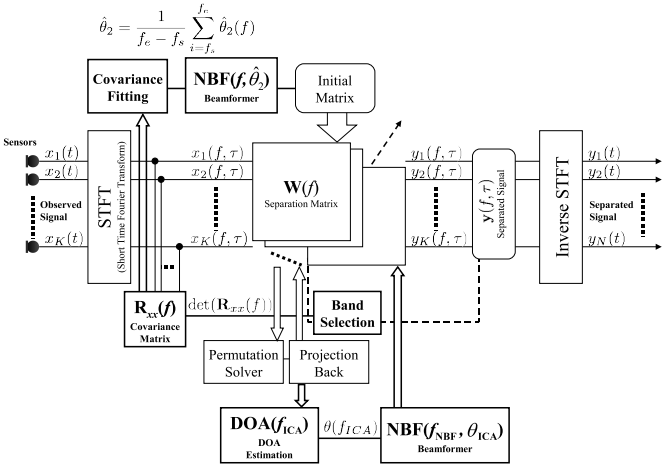

Figure 2: *Block diagram of* tinyICA.

Table 1: *ICA parameters*

FFT Size	1024 [sample]
FFT Shift	512 [sample]
Learning Time	3 [sec]
Iteration	10,20,40,80,160,240,320,400,500 [times]
Step Size	0.01
Perm. Solver	DOA

3.5. Presumption of the amount of computation

We estimate the number of operations (multiplication, addition as floating operations) and memory size to evaluate suitability for *tiny* DSP modules. They are shown in Table 2.

The parameters of ICA are shown in Table 1. The differences of the estimation between conventional method and the proposed method are 1) the number of iterations, 2) the matrix initialization step and 3) the number of selected bands. For conventional FDICA, 1) 500 iterations, 2) the identity matrix and 3) full selected bands are used. For the proposed method, 1) 100 iterations, 2) the proposed matrix initialization step and 3) 50 selected bands are used.

From Table 2, the proposed method shows improvement of 15% calculations and 17% memory size compared to conventional FDICA. In other words, a reduction of amount of computation is over 80%.

4. Experimental Results

In this section, we show the performance of *tinyICA* by convergence and separation properties by the source separation simulation.

Signals for simulation are shown in Table 3, and they are recorded in an anechoic room and a reverberant room. The recording condition is shown in Fig-3: the each voice signal is played back through a loudspeaker and recorded individually. The signals are mixed, when they are simulated. The parameters of ICA and the microphone array are shown in Table 1 and Table 4. Moreover, we use Noise Reduction Rate (NRR) [7] which indicates a difference between an output (processed) SNR and an input SNR.

For the separation matrix initialization, NRR is evaluated in terms of the number of iterations to measure a convergence property, and in addition, based on a position symmetry, we only simulate a quarter round position. We change the separation matrix initialization step: the proposed method, the identity matrix and the *ideal* null beamformer (assume that the directions of sources are known). In Fig-4 and Fig-5, the proposed method shows nearly ideal performance corresponding to the

Table 2: *Amounts of computation*

	conventional	proposed
number of operations [MOPs]	146	22
size of memory [Bytes]	1156k	198k

Table 3: *Signals for simulation*

	Anecho.	Reverb.
Samp. Freq.	8 [kHz]	
Rev. Time	–	400 [msec]
2 Voices	Male(2), Female(2), *known* : 0 [deg] *unknown* : -90,-45 [deg]	same as left

Table 4: *Array parameters*

Microphone	OMNI, SHURE SM93
Num. of Mic.	2
Interval of Mic.	3.6 [cm]

ideal null beamformer. It shows that the faster convergence and almost 100 iterations is enough. In addition, the similar trend can be seen between the anechoic condition and the reverberant condition. Moreover, if the position of the *unknown* source is changed, the trend of the performance of the proposed method is preserved. These aspects imply the robustness of the proposed method.

For the learning band selection, NRR is evaluated in terms of the number of selected bands. We use 100 iterations, the proposed separation matrix initialization method step and the sources positioned 0 and -90 degree to evaluate performance. In anechoic and 50 bands selection, NRR indicates around 15 [dB] in Fig-6. In reverberant, there are less NRR depressions around 1–2[dB]. These results indicate on the performance is practical enough, and we get a foothold to implement ICA on *tiny* DSP modules.

As a consequence, it is shown that the integrated method has the faster convergence, less amount of computation, and practical performance.

5. Conclusions

The semi-blind source separation method with the less amount of computation —*tinyICA*— is introduced.

It is shown that ICA is better initialized by covariance fitting, and it contributes to the faster convergence and less amount of computation. The learning band selection achieves a significant reduction of computation with the practical separation performance using acoustical sense of the separation matrix. Finally, a total reduction of amount of computation is over 80%.

In the future, our plan is to evaluate the proposed method performance in actual environments. In addition, we are planning to report a theoretical and experimental verification for the other scalar values for learning band selection.

6. Acknowledgements

We would like to express our gratitude to Dr. Hiroshi SARUWATARI and Mr. Yu TAKAHASHI at Nara Institute of Technology in Japan for valuable advice. We thank our colleages for discussions and helps.

7. References

[1] P. Smaragdis. Blind separation of convolved mixtures in the frequency domain. *Neurocomputing*, 22:21–34, 1998.

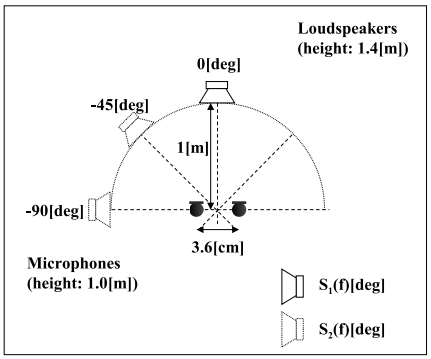

Figure 3: *Recording conditions*

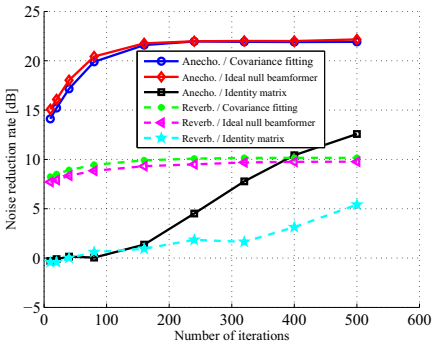

Figure 4: *NRR for different matrix initialization (0/-90[deg])*

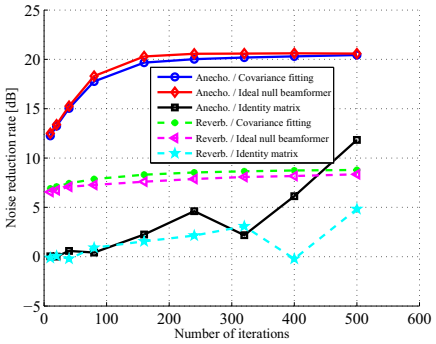

Figure 5: *NRR for different matrix initialization (0/-45[deg])*

[2] T. Hiekata and et.al. Development and evaluation of pocket-size blind source separation microphone. *Proc. of IWAENC08*, Sep 2008.

[3] P. Stoica and et.al. Robust capon beamforming. *IEEE Signal Processing Letters*, 10(6):172–175, Jun 2003.

[4] N. Murata and et.al. An on-line algorithm for blind source separation on speech signals. *Proc. of NOLTA1998*, 3:923–926, 1998.

[5] S. Kurita and et.al. Evaluation of blind signal separation method using directivity pattern under reverberant conditions. *Proc. of ICASSP2000*, 5:3140–3143, 2000.

[6] S. Araki and et.al. The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech. *IEEE Trans. SAP*, 11(2):109–116, mar 2003.

[7] H. Saruwatari and et.al. Blind source separation combining independent component analysis and beamforming. *EURASIP Jouranl on Applied Signal Processing*, 2003(1):1135–1146, jan 2003.

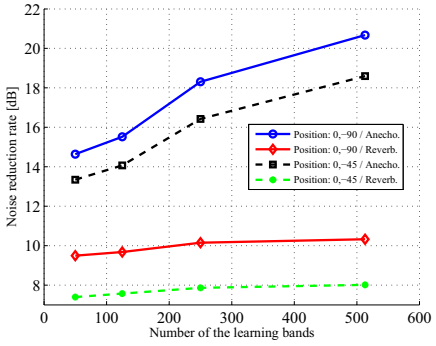

Figure 6: *NRR for earning band selection*