ONLINE IVA WITH ADAPTIVE LEARNING FOR SPEECH SEPARATION USING VARIOUS SOURCE PRIORS

Suleiman Erateb[*], Mohsen Naqvi[†] and Jonathon Chambers[†]
[*]Wolfson School of Mechanical, Manufacturing and Electrical Engineering,
Loughborough University, LE11 3TU, UK
[†]School of Electrical and Electronic Engineering, Newcastle University, NE1 7RU, UK
[*]s.erateb@lboro.ac.uk, [†]{mohsen.naqvi, jonathon.chambers}@newcastle.ac.uk

Abstract

Independent vector analysis (IVA) is a frequency domain blind source separation (FDBSS) technique that has proven efficient in separating independent speech signals from their convolutive mixtures. In particular, it addresses the problematic permutation problem by using a multivariate source prior. The multivariate source prior models statistical inter dependency across the frequency bins of each source and the performance of the method is dependent upon the choice of source prior. The online form of the IVA is suitable for practical real time systems. Previous online algorithms use a learning rate that does not introduce a robust way to control the learning as a function of the proximity to the target solution. In this work, we propose a new adaptive learning scheme to improve the convergence speed and steady state separation performance. The speech signals are modelled by two different source priors; a super-Gaussian distribution and a generalized Gaussian distribution. The experimental results confirm improved performance with real room impulse responses and real recorded speech signals.

Index Terms -- Blind source separation, convolutive mixture, independent vector analysis, online, adaptive learning, room impulse responses

I. INTRODUCTION

BSS is generally a statistical signal processing approach to solve the cocktail party problem (CPP). CPP describes the problem of separating different sounds in a cocktail party environment [1]. BSS is concerned with extracting source signals from their observed mixtures without information about the sources and the mixing process. The observed signals are obtained at a set of spatially distinct sensors, each receiving a different combination of the source signals. The mixing process becomes convolutive due to reverberations in the real room environment. Separation may be achieved in different ways according to the amount of prior information available [2]. Time domain methods are generally not appropriate for the convolutive BSS (CBSS) problem due to the computational complexity [3]. In order to reduce the computational cost, frequency domain methods have been proposed to solve the CBSS problem. The convolution operation in the time domain becomes multiplication in the frequency domain [4].

Independent vector analysis (IVA) is a method to tackle BSS in the frequency domain. The technique has proven efficient in separating independent speech signals from convolutive mixtures [5]. It solves, algorithmically, the problematic permutation problem inherent in independent component analysis (ICA) [6]. IVA extends ICA from a univariate source signal model to a multivariate one. The multivariate source prior models statistical inter dependency across the frequency bins of each source.

The original IVA method proposed in [5] runs in an offline batch manner where the entire set of input samples is gathered before calculating the parameters. This approach is not applicable to practical online systems. A block-based approach can be applied to implement a real time BSS system [7]. However, this approach encompasses heavy computational load. A fully online version of the IVA algorithm was proposed in [8] which is suitable for practical embedded systems. In online IVA, the coefficients of the separation filter are updated at every time frame.

Usual online IVA methods use a fixed learning rate to update the unmixing matrix. If the learning rate is set to a high value, the solution convergences faster with large fluctuations. For small learning rate value, the convergence is slower with smoother solution. In this paper, the contribution is to introduce a new adaptive learning scheme to improve the performance in terms of convergence time and steady state separation. The scheme combines the advantages of the high and small values of the learning rate. The learning rate is controlled by a Frobenius norm as a measure of the proximity to the target solution, which is extracted from the learning gradient adopted. Two source priors are used to model the speech signals; the super-Gaussian distribution proposed in the original IVA [5] based on a spherically symmetric Laplace (SSL) distribution and a generalized Gaussian distribution proposed in [9] which exploits fourth order inter-frequency correlation and was previously only tested on the batch IVA.

The original IVA algorithm was evaluated using synthetic room impulse responses (RIRs) based on the image source method (ISM) [10] which are artificial and do not represent a real life room environment. In this work, the proposed scheme is evaluated on real room impulse responses which are termed as binaural room impulse responses (BRIRs) [11]. Real recorded speech signals from the TIMIT acoustic-phonetic continuous speech corpus [12] are used as the source signals.

978-1-5386-1663-5/17/$31.00 ©2017 IEEE

This paper is organized as follows: in Section II, the online IVA method is introduced, the adopted source priors are described and the proposed scheme is presented. The simulations and experimental results are shown in Section III. Finally, conclusions are drawn and future work is discussed in Section IV.

II. BACKGROUND THEORY

The BSS problem can be concisely stated as the estimation of *N* source signals from *M* observed mixture signals that are unknown functions of the sources.

A. Online IVA

The noise free FDCBSS online IVA mixing and separation models are described as [8]:

$$x_j^{(k)}[n] = \sum_{i=1}^{N} h_{ji}^{(k)}[n] s_i^{(k)}[n] \tag{1}$$

$$\hat{s}_i^{(k)}[n] = \sum_{j=1}^{M} w_{ij}^{(k)}[n] x_j^{(k)}[n] \tag{2}$$

where $x_j^{(k)}[n]$, $s_i^{(k)}[n]$ and $\hat{s}_i^{(k)}[n]$ are the *j*-th observation value, the *i*-th source signal and *i*-th estimated source at time frame *n* at the *k*-th frequency bin respectively. $h_{ji}^{(k)}[n]$ and $w_{ij}^{(k)}[n]$ are the mixing and unmixing filter coefficients at time frame *n* at the *k*-th frequency bin respectively. $k = 1, 2, \dots, K$, and *K* is the number of frequency bins.

IVA uses a multivariate source prior to retain the dependency between different frequency bins of each source. The independence is measured by the Kullback-Leibler (KL) divergence between the exact joint probability density function of the estimated source vectors $p(\hat{\boldsymbol{s}}_1, \dots, \hat{\boldsymbol{s}}_N)$ and the product of marginal probability density functions of the individual source vectors $\prod_{i=1}^{N} q(\hat{\boldsymbol{s}}_i)$ [4]:

$$C = \mathcal{KL}\left(p(\hat{\boldsymbol{s}}_1, \dots, \hat{\boldsymbol{s}}_N) \middle\| \prod_1^N q(\hat{\boldsymbol{s}}_i)\right) \tag{3}$$

The source prior $q(\hat{s}_i)$ in the cost function is a vector across all frequency bins. Each source is multivariate and the KL divergence would be minimized when the dependency between the source vectors is removed but the inherent dependency between the components of each vector is preserved. The learning algorithm for the parameters of the separating filters is derived by minimizing the KL cost function using a gradient descent method [13].

B. Source Priors

The performance of the IVA algorithm greatly depends on the multivariate inter-dependency model used as a source prior. The IVA algorithm proposed in [5] defines the source prior as a dependent multivariate super-Gaussian distribution in the form:

$$q(\boldsymbol{s}_i) = \alpha exp\left(-\sqrt{(\boldsymbol{s}_i - \boldsymbol{\mu}_i)^H \Sigma_i^{-1} (\boldsymbol{s}_i - \boldsymbol{\mu}_i)}\right) \tag{4}$$

where $(\cdot)^H$ denotes Hermitian transpose, $\boldsymbol{\mu}_i$ and Σ_i are the mean vector and covariance matrix of the *i*-th source signal, respectively. Assuming zero mean, identity covariance matrix and unity standard deviation:

$$q(\boldsymbol{s}_i) = \alpha\, exp\left(-\sqrt{\sum_{k=1}^{K} \left|s_i^{(k)}\right|^2}\right) \tag{5}$$

The resulting non-linear multivariate score function vector is given as:

$$\varphi^{(k)}\left(\hat{s}_i^{(1)} \dots \hat{s}_i^{(k)}\right) = \frac{\hat{s}_i^{(k)}}{\sqrt{\sum_{k=1}^{K} \left|\hat{s}_i^{(k)}\right|^2}} \tag{6}$$

Y. Liang et al. [9] proposed a generalized Gaussian source prior in the form:

$$q(\boldsymbol{s}_i) = \alpha exp\left(-\sqrt[3]{(\boldsymbol{s}_i - \boldsymbol{\mu}_i)^H \Sigma_i^{-1} (\boldsymbol{s}_i - \boldsymbol{\mu}_i)}\right) \tag{7}$$

The distribution has heavier tails than the distribution in the original IVA and the authors claim it is more robust to outliers present in statistically non-stationary speech. Assuming zero mean, identity covariance matrix and unity standard deviation:

$$q(\boldsymbol{s}_i) = \alpha\, exp\left(-\sqrt[3]{\sum_{k=1}^{K} \left|s_i^{(k)}\right|^2}\right) \tag{8}$$

The resulting non-linear multivariate score function vector is given as:

$$\varphi^{(k)}\left(\hat{s}_i^{(1)} \dots \hat{s}_i^{(k)}\right) = \frac{\hat{s}_i^{(k)}}{\sqrt[3]{\left(\sum_{k=1}^{K} \left|\hat{s}_i^{(k)}\right|^2\right)^2}} \tag{9}$$

C. The Online Learning Algorithm

The coefficients of the separation filter coefficients are updated at every time block using a normalized learning rate as follows [8]:

$$w_{ij}^{(k)}[n+1] = w_{ij}^{(k)}[n] + \eta \sqrt{(\xi^{(k)}[n])^{-1}} \Delta w_{ij}^{(k)}[n] \tag{10}$$

where η is the learning rate and $\xi^{(k)}[n]$ is a normalisation factor given by:

$$\xi^{(k)}[n] = \beta\xi^{(k)}[n-1] + (1-\beta)\sum_{i=1}^{N}\left|x_i^{(k)}[n]\right|^2/N \quad (11)$$

where $\beta \in [0,1]$ is a smoothing factor and $\Delta w_{ij}^{(k)}[n]$ is the gradient with nonholomonic constraint of the current frame as follows:

$$\Delta w_{ij}^{(k)}[n] = \sum_{l=1}^{N}\left(\Lambda_{ij}^{(k)}[n] - \Re_{ij}^{(k)}[n]\right) w_{ij}^{(k)}[n] \quad (12)$$

where $\Lambda^{(k)}[n]$ is a diagonal matrix based on the non-linear score function ($\Lambda_{ii}^{(k)}[n] = \Re_{ii}^{(k)}$ and $\Lambda_{il}^{(k)}[n] = 0$ when $i \neq l$) and $\Re^{(k)}[n]$ is the online scored correlation matrix at the current frame termed as:

$$\Re_{ij}^{(k)}[n] = \varphi^{(k)}\left(\hat{s}_i^{(1)} \dots \hat{s}_i^{(k)}\right)\hat{s}_l^{(k)*}[n] \quad (13)$$

and $(\cdot)^*$ denotes the conjugate operator.

D. New Adaptive Online Learning

The gradient $\Delta w_{ij}^{(k)}[n]$ converges to zero as $\Lambda_{ij}^{(k)}[n]$ approaches $\Re_{ii}^{(k)}[n]$ i.e. $\Lambda_{ij}^{(k)}[n] - \Re_{ij}^{(k)}[n]$ approaches zero. We therefore assign:

$$G^{(k)}[n] = \left\|\Lambda^{(k)}[n] - \Re^{(k)}[n]\right\|_F \quad (14)$$

where $\|.\|_F$ denotes the Frobenius norm.

We utilise the descending behaviour of $G^{(k)}[n]$ as a gear-shifting type operator. In the initial stages the learning rate is set to a high value to move faster towards the solution. Then it decreases as the system converges to reduce the fluctuations and improve stability. We define a new normalised learning rate at time frame *n* as:

$$\eta^{(k)}[n] = \eta_0 \frac{\left\|G^{(k)}[n]\right\|_F}{\|G^{(k)}[1]\|_F} \quad (15)$$

where η_0 is the initial learning rate. In a non-stationary environment $G^{(k)}[1]$ could be reinitialized. Then $\eta^{(k)}[n]$ is smoothed using as follows:

$$\eta^{(k)}[n] = \left[\lambda\eta^{(k)}[n-1] + (1-\lambda)\eta^{(k)}[n]\right] \quad (16)$$

where $\lambda = 0.99$ is an empirically determined smoothing factor. $\eta^{(k)}[n]$ will start with the initial value η_0 for the first frame and then it decreases as *n* increases. The online update equation is adjusted accordingly as:

$$w_{ij}^{(k)}[n+1] = w_{ij}^{(k)}[n] + \eta^{(k)}[n]\sqrt{(\xi^{(k)}[n])^{-1}}\Delta w_{ij}^{(k)}[n] \quad (17)$$

III. SIMULATION STUDIES

In this section we evaluate the proposed adaptive learning scheme for the IVA method using the two source priors.

A. Experimental setup

Experiments were conducted to evaluate the performance of the proposed scheme using both source priors. For our study, a two-input (speaker) two-output (microphone) (TITO) system under spatially stationary conditions was adopted. The proposed scheme was evaluated on real room impulse responses [11]. Real recorded speech signals, from the TIMIT acoustic-phonetic continuous speech corpus [12], were used as the source signals. The sources were convolved with the room impulse response to generate the mixture signals at the microphones. The signal to distortion ratio (SDR) was used to measure the separation performance by using the SISEC toolbox [14]. SDR is defined by the power ratio between the components related to the target source and interference sources plus artifacts from the separation algorithm:

$$SDR = 10\log_{10}\frac{\left\|\boldsymbol{s}_{target}\right\|^2}{\left\|\boldsymbol{e}_{interf} + \boldsymbol{e}_{artif}\right\|^2} \quad (18)$$

where $\|.\|^2$ denotes the energy of the signal, $\boldsymbol{s}_{target}$ is the source of interest, $\boldsymbol{e}_{interf}$ is the interference introduced by the other sources and $\boldsymbol{e}_{artif}$ is the artifacts error term. SDR is directly proportional to the quality of source separation.

The room impulse responses were obtained from the BRIRs database [11] which was recorded using a dummy head to simulate the effect of a human head in a real acoustic environment. BRIRs were measured for 21 different relative source locations, consisting of all combinations of seven source azimuths (0°, 15°, 30°, 45°, 60°, 75°, and 90°) and three source distances (0.15m, 0.40m, and 1m) from the centre point between the ears of the head.

The room layout and experimental setup are illustrated in Figure 1. The microphones were placed at the centre of the room. The sources were placed at 0.40m from the centre of the microphones. Source s_1 was at a fixed position perpendicular to both microphones at 0° and source s_2 was at five different angles (15° to angle 75°) relative to source s_1. The inter-microphone distance is 15cm. The different experiment parameters used for our simulations are shown in Table 1.

For both learning algorithms the leaning rate η was set to a value that makes the system converge quickly whilst maintaining stability for all source angles. Such large value of η makes the algorithm converge faster but produces high fluctuations in the steady state, which may lead to instability. η was set to 0.5 and η_0 to 2.0.

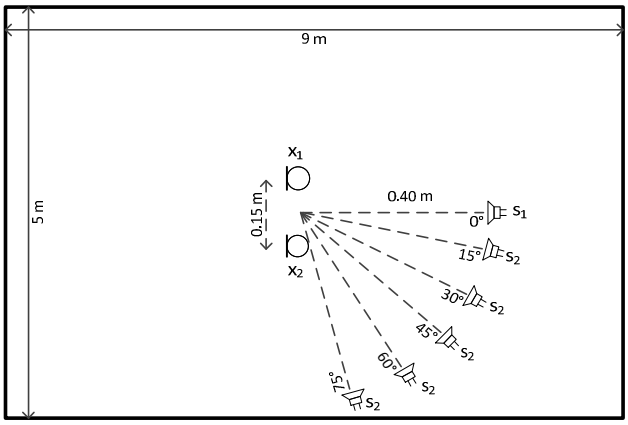

Figure 1. Room layout and experimental setup

Table 1 Experiment Parameters

Room dimensions	$9m \times 5m \times 3.5m$
Inter-Microphone distance	0.15 *m*
Source distance	0.40 *m*
Source 1 position	0°
Source 2 positions	15°, 30°, 45°, 60°, 75°
The length of the DFT	2048
Sampling frequency	8 *kHz*
Window type	Hanning
Sound propagation speed	343 *m/s*
Reverberation time	565 *ms*
η for original method	0.5
η_0 for proposed method	2.0
Smoothing factor β	0.5

B. Results

Ten speech signal pairs were randomly selected from the TIMIT database to evaluate the algorithms. The sources were separated from the generated mixtures and the calculated SDR was averaged over the ten results. Figure 2 shows the SDR convergence plots for the different algorithms with source s_2 at angles 30° and 60°over a period of 150 seconds.

We evaluated the separation performance in terms of the convergence time and the steady state SDR. The steady state is considered to be the average SDR of the last 50 seconds and we define the convergence time as the time it takes the algorithm to reach 80% of the final steady state SDR. The values of the convergence times in seconds are shown in Table 2 and the values of the average steady state SDR in dB are shown in Table 3.

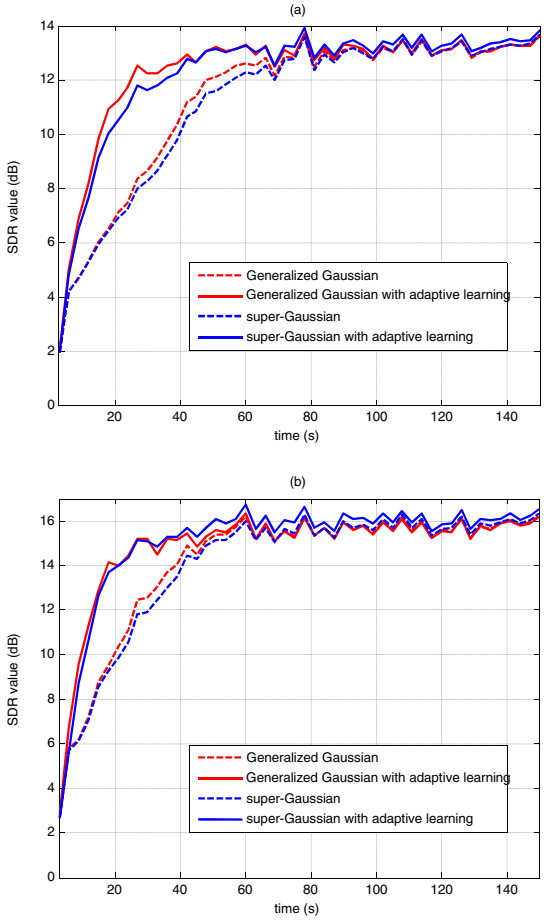

Figure 2. SDR convergence in (dB) for the various algorithms averaged over 10 speech mixtures, where (a) s_2 is at angle 30° and (b) s_2 is at angle 60°

Table 2. Convergence time in seconds for various algorithms at different source s_2 positions

	Angle				
Source Prior	**15°**	**30°**	**45°**	**60°**	**75°**
super-Gaussian [5]	75	42	38	35	31
generalized Gaussian [9]	75	40	35	30	25
[5] with Adaptive learning	50	22	17	16	14
[9] with Adaptive learning	40	17	15	14	14

Table 3. Average steady state SDR in (dB) for various algorithms at different source s_2 positions

	Angle				
Source Prior	**15°**	**30°**	**45°**	**60°**	**75°**
super-Gaussian [5]	9.25	13.24	14.94	15.82	16.36
generalized Gaussian [9]	9.18	13.18	14.85	15.71	16.22
[5] with Adaptive learning	9.26	13.37	15.11	16	16.56
[9] with Adaptive learning	9.22	13.2	14.88	15.73	16.25

The results show, generally, the larger the angle of source s_2 the better the performance in both the convergence speed and steady state SDR, with less disparity as the angle increases. The results exhibit a consistent and considerable improved performance of the proposed scheme in terms of the convergence speed. It reduces the convergence time by approximately an average of 20.5 seconds (46%) using the super-Gaussian source prior and by an average of 21 seconds (51%) using the generalized Gaussian source prior. The proposed scheme with the generalized Gaussian source prior [9] converges faster than with the super-Gaussian source prior [5]. The former is faster, on average, by 3.8 seconds (16%).

In terms of the steady state all algorithms converge to an SDR value with small variations. This demonstrates the success of the adaptive learning scheme in reducing the learning rate as the algorithm convergence to the target solution. The average steady state SDR improvements are approximately 0.15 dB and 0.05 dB using the super-Gaussian source prior the generalized Gaussian source prior respectively. The proposed scheme with the super Gaussian source prior [5] achieves better separation performs than with the super-Gaussian source prior [9] by approximately 0.2 dB.

IV. CONCLUSION

In this paper, an adaptive learning based scheme to control the learning rate has been proposed in order to improve the performance and convergence properties of the online IVA algorithm. The scheme was tested and compared with the original IVA algorithm using real room impulse responses and real recordings. The experimental results have shown the new scheme yields faster and smoother convergence time, better separation performance measured by SDR. On the balance of results, we believe the best overall performance is achieved with adaptive learning using the generalized Gaussian source prior. The scheme incurs an additional computational cost calculating the normalised Frobenius norm at every time frame. Future work will include evaluating the new scheme exploring other source prior distributions such as the Student's t distribution [15, 16] and mixed source prior [17]. An interesting research point will be combining the super Gaussian and the generalized Gaussian source priors to acquire the best aspect of each distribution.

REFERENCES

[1] E. C. Cherry and W. K. Taylor. Some further experiments upon the recognition of speech, with one and with two ears. *J. Acoust. Soc. Am,* vol. 26, pp. 554-559, 1954.

[2] N. Das, A. Routray, P. K. Dash and D. India, "ICA methods for blind source separation of instantaneous mixtures: A case study," *Neural Information Process. Letters and Reviews,* vol. 11, pp. 225-246, 2007.

[3] M. S. Pedersen, J. Larsen, U. Kjems and L. C. Parra, "A survey of convolutive blind source separation methods," *Multichannel Speech Processing Handbook,* pp. 1065-1084, 2007.

[4] L. Parra and C. Spence, "Convolutive blind separation of non-stationary sources," *Speech and Audio Processing, IEEE Transactions on,* vol. 8, pp. 320-327, 2000.

[5] T. Kim, H. T. Attias, S. Lee and T. Lee, "Blind source separation exploiting higher-order frequency dependencies," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 15, pp. 70-79, 2007.

[6] E. Bingham and A. Hyvärinen, "A fast fixed-point algorithm for independent component analysis of complex valued signals," *Int. J. Neural Syst.,* vol. 10, pp. 1-8, 2000.

[7] R. Mukai, H. Sawada, S. Araki and S. Makino, "Blind Source Separation for moving speech Signals using blockwise ICA and residual crosstalk subtraction," *IEICE Trans. Fundam*., vol. E87-A, no. 8, pp. 530-538, 2004.

[8] T. Kim, "Real-time independent vector analysis for convolutive blind source separation," *Circuits and Systems I: Regular Papers, IEEE Transactions on*, *57*(7), pp.1431-1438, 2010.

[9] Y. Liang, S. M. Naqvi, and J. A. Chambers. "Independent vector analysis with a multivariate generalized Gaussian source prior for frequency domain blind source separation," *Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on*, pp. 6088-6092, 2013.

[10] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, pp. 943–950, 1979.

[11] B. Shinn-Cunningham, N. Kopco, and T. Martin, "Localizing nearby sound sources in aclassroom: Binaural room impulse responses," *J. Acoust. Soc. Amer.*, vol. 117, pp. 3100–3115, 2005.

[12] J. S. Garofolo et al, "DARPA TIMIT acoustic phonetic continuous speech corpus CDROM," *NASA STI/Recon technical report n*, 1993.

[13] S. I. Amari, A. Cichocki, and H. H. Yang, "A new learning algorithm for blind signal separation," *Adv. Neural Inf. Process. Syst.*, vol. 8, pp. 752–763, 1996.

[14] E. Vincent, R. Gribonval and C. Févotte, "Performance measurement in blind audio source separation," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 14, pp. 1462-1469, 2006.

[15] Y. Liang, G. Chen, S.M.R. Naqvi and J.A Chambers, "Independent vector analysis with multivariate Student's t distribution source prior for speech separation," *Electronics Letters*, vol. 49, pp. 1035-1036, 2013.

[16] J. Harris et al, "Real-time independent vector analysis with Student's t source prior for convolutive speech mixtures," *Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on*, pp. 1856-1860, 2015.

[17] W. Rafique, S. Erateb, S. M. Naqvi, S. S. Dlay, and J. A. Chambers. "Independent vector analysis for source separation using an energy driven mixed Student's t and super Gaussian source prior," *Signal Processing Conference (EUSIPCO), 24th European*, pp. 858-862, 2016.