STAIRS 2014
U. Endriss and J. Leite (Eds.)

doi:10.3233/978-1-61499-421-3-181

Supervised Separation of Speech from Background Piano Music using a Nonnegative Matrix Factorization Approach

A. MARTINEZ-COLÓN [a,1], F. J. CANADAS-QUESADA [a] and P. VERA-CANDEAS [a] and N. RUIZ-REYES [a] and F. MORENO-FUENTES [a]
[a] *Telecommunication Engineering Department, University of Jaén, Spain*

Abstract. This paper presents a supervised algorithm for separating speech from background non-stationary noise (piano music) in single-channel recordings. The proposed algorithm, based on a nonnegative matrix factorization (NMF) approach, is able to extract speech sounds from isolated or chords piano sounds learning the set of spectral patterns generated by independent syllables and piano notes. Moroever, a sparsity constraint is used to improve the quality of the separated signals. Our proposal was tested using several audio mixtures composed of real-world piano recordings and Spanish speech showing promising results.

Keywords. Sound separation, Non-negative matrix factorization, training, supervised, sparse, interference

1. INTRODUCTION

Separation of a target source (speech) from background non-stationary noise (piano) is still a challenging problem in artificial intelligence, signal processing and music research. The speech refers to vocal sounds used in a human communication whereas the piano sound refers to the sounds generated by a piano instrument.

Several approaches to separate speech and background non-stationary noise have been proposed in the last years [1] [2] [3]. Schmidt et.al [1] presented a method, based on non-negative sparse coding, for reducing wind noise in recordings of speech based on a pre-estimated source model only for the noise. In [2], a sparse latent variable model is proposed which can be employed for the decomposition of time/frequency distributions to perform separation of sources from single-channel recordings. In [3], speech is modeled using a non-negative hidden Markov model, which uses multiple non-negative dictionaries and a Markov chain to jointly model spectral structure and temporal dynamics of speech.

Non-negative matrix factorization (NMF) has been successfully applied in the field of speech and music processing in recent years [4] [5] [6] [7] [8] [9]. Lee and Seung

[1]Corresponding Author: A. Martinez-Colón, Telecommunication Engineering Department, University of Jaén, Spain ; E-mail: fcanadas@ujaen.es

[10] [11] developed standard NMF, a technique for multivariate data analysis in which an input magnitude spectrogram, represented by a matrix X, is decomposed into the product of two non-negative matrices W and H,

$$X \approx WH \tag{1}$$

where each column of the basis matrix W represents a spectral pattern from an active sound source. Each row of the gains matrix H represents the time-varying activations of a spectral pattern factorized in the basis matrix. In general, NMF approaches can be classified into three categories [12]

Supervised: all spectral patterns both the target and non-target source are trained previously to the separation stage.

Semisupervised: only spectral patterns from the target source or non-target source are trained previously to the separation stage.

Unsupervised: no training stage is used. Instead, the factorization process is performed using different type of constraints.

In this work, we propose a supervised NMF approach to separate speech and polyphonic piano music in single-channel recordings. Our proposal is composed of two stages: training and separation. In the training stage, the system learn the spectral patterns from sounds related to syllables of Spanish speech and sounds from musical isolated piano notes. Using the previous patterns, our proposed algorithm is able to decompose a monaural audio mixture into speech and piano signals. As it will be explained later, we have used a sparsity constraint in order to improve the quality of the speech and minimizing the interference of the piano and vice versa.

This paper is organized as follows. In section 2, the proposed method is depicted in detail. In section 3, test data, experimental setup and metrics are explained. In section 4, experimental results are shown. Finally, the conclusions and future work are presented in section 5.

2. PROPOSED METHOD

The scheme of the proposed method is shown in Figure 1. Because of our proposed method is based on a supervised NMF approach, it needs a two training stages. The first one is related to factorize the spectral patterns of the syllables of the speech. The second one is related to factorize the spectral patterns of the piano notes. The most used cost functions are the Euclidean (EUC) distance, the generalised Kullback-Leibler (KL) and the Itakura-Saito (IS) divergences. However, in this work, the KL and IS divergences have been analyzed because they have provided the best results in the separation stage.

2.1. Speech training stage

To obtain the spectral patterns W_s of the speech, a speech database D_s was generated recording, using a portable recorder *Zoom H4n* [13], a set of different syllables of the Spanish language. Specifically, the speech database is composed of 420 syllables: 5 syllables of one letter, 118 syllables of two letters, 291 syllables of three letters and 6 sylla-

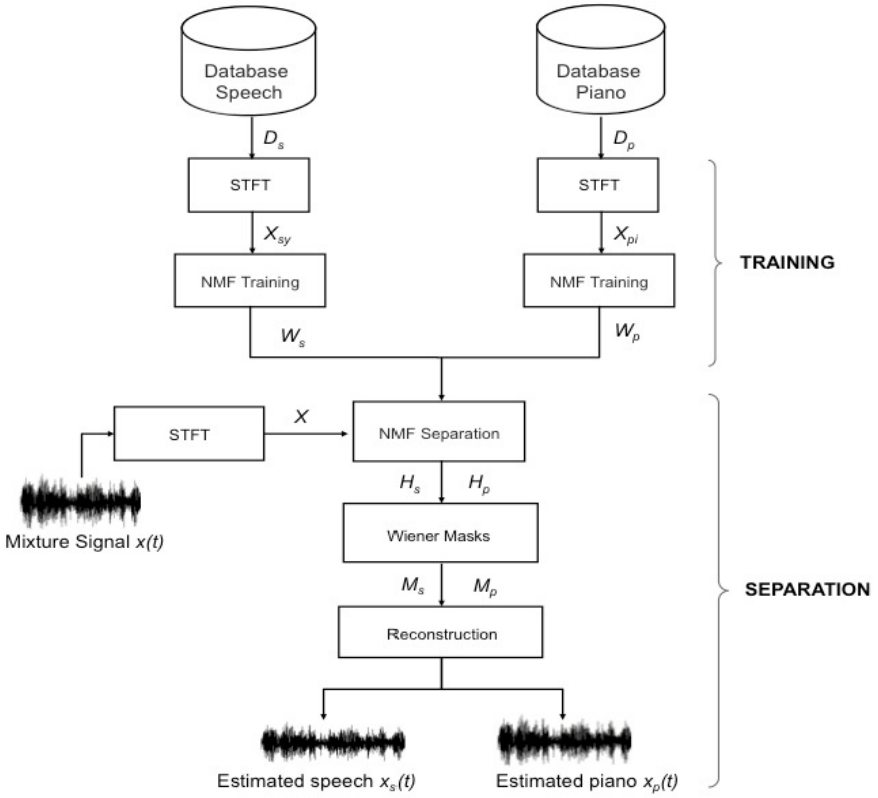

Figure 1. Overview of the proposed supervised NMF approach

bles of four letters. The selection of the syllables was made taking into account the most likely syllables to be spoken in the Spanish language. In the factorization process, we have considered K_s spectral patterns to model each syllable.

In order to estimate the speech basis W_s or gains H_s matrices, the iterative algorithm proposed in [11] [12] can be applied,

- *Kullback-Leibler* divergence

$$W_s = W_s \odot \frac{\left(X_{sy} \odot (W_s \cdot H_s)^{-1}\right) \cdot H_s^T}{\mathbf{1} \cdot H_s^T} \tag{2}$$

$$H_s = H_s \odot \frac{W_s^T \cdot (X_{sy} \odot (W_s \cdot H_s)^{-1})}{W_s^T \cdot \mathbf{1}} \tag{3}$$

- *Itakura-Saito* divergence

$$W_s = W_s \odot \frac{\left(X_{sy} \odot (W_s \cdot H_s)^{-2}\right) \cdot H_s^T}{(W_s \cdot H_s)^{-1} \cdot H_s^T} \tag{4}$$

$$H_s = H_s \odot \frac{W_s^T \cdot \left(X_{sy} \odot (W_s \cdot H_s)^{-2}\right)}{W_s^T \cdot (W_s \cdot H_s)^{-1}} \tag{5}$$

where $\odot$ is the element-wise product operator, ^{T} is the transpose operator, X_{sy} is the magnitude spectrogram of each syllable and $\mathbf{1}$ is an all-one elements matrix. The speech training procedure is summarized in Algorithm 1

Algorithm 1 Training of Speech Spectral Patterns

1 **for** each syllable **do**
2 Compute the speech magnitude spectrogram X_{sy} from a syllable of the database D_s.
3 Initialise all rows of the gain matrix H_s with random positive values.
4 Initialise all columns of the basis matrix W_s with random positive values.
5 Update bases W_s using eq. (2) or (4)
6 Update gains H_s using eq. (3) or (5)
7 Repeat steps 5-6 until the algorithm converges (or the maximum number of iterations $MaxIter$ is reached).
8 **end for**

2.2. Piano training stage

To obtain each spectral patterns W_p of a piano instrument, the piano database D_p was generated using samples of notes from a piano instrument [14]. Specifically, the piano database is composed of 88 sounds of isolated piano notes played on a normal intensity. In the factorization process, we have considered K_p spectral patterns to model each piano note.

The piano update rules to compute W_p and H_p are similar to speech ones (see eq. (2-5)) replacing X_{sy} for the magnitude spectrogram X_{pi} of each musical note from a piano instrument and replacing W_s to W_p and H_s to H_p. The piano training procedure is summarized in Algorithm 2

Algorithm 2 Training of Piano Spectral Patterns

1 **for** each note **do**
2 Compute the piano magnitude spectrogram X_{pi} from a piano note of the database D_p.
3 Initialise all rows of the gain matrix H_p with random positive values.
4 Initialise all columns of the basis matrix W_p with random positive values.
5 Update bases W_p using eq. (2) or (4)
6 Update gains H_p using eq. (3) or (5)
7 Repeat steps 5-6 until the algorithm converges (or the maximum number of iterations $MaxIter$ is reached).
8 **end for**

As a consequence of using a supervised NMF approach, W_s and W_p are pre-computed and known in the training stages and held fixed during the factorization process in the separation stage.

2.3. Separation stage

The magnitude spectrogram X of a mixture signal $x(t)$ can be performed by the Short-Time Fourier Transform (STFT) using a N samples *Hamming* window and J samples time shift. The mixture spectrogram X is composed of a speech X_s and a piano X_p spectrograms,

$$X = X_s + X_p \tag{6}$$

, where each spectrogram X_s or X_p represents the specific spectral features exhibited by the speech and piano instrument. In this manner, our factorization model is defined

$$\hat{X} \approx \hat{X}_s + \hat{X}_p \approx (W_s * H_s) + (W_p * H_p) \tag{7}$$

being $\hat{X}, \hat{X}_s, \hat{X}_p, W_s, W_p, H_s$ and H_p the estimated mixture spectrogram, the estimated speech spectrogram, the estimated piano spectrogram, the speech and piano spectral patterns and the speech and piano gains.

The speech H_s and piano H_p gains update rules are shown using the *Kullback-Liebler* divergence (eq. (8) and (9)) and *Itakura-Saito* divergence (eq. (10) and (11)) [12] with a sparsity (speech λ_s or piano λ_p) constraint [4]

$$H_s = H_s \odot \frac{W_s^T \cdot (X \odot ((W_s \cdot H_s) + (W_p \cdot H_p))^{-1})}{W_s^T \cdot 1 + \lambda_s} \tag{8}$$

$$H_p = H_p \odot \frac{W_p^T \cdot (X \odot ((W_s \cdot H_s) + (W_p \cdot H_p))^{-1})}{W_p^T \cdot 1 + \lambda_p} \tag{9}$$

$$H_s = H_s \odot \frac{W_s^T \cdot (X \odot ((W_s \cdot H_s) + (W_p \cdot H_p))^{-2})}{W_s^T \cdot ((W_s \cdot H_s) + (W_p \cdot H_p))^{-1} + \lambda_s} \tag{10}$$

$$H_p = H_p \odot \frac{W_p^T \cdot (X \odot ((W_s \cdot H_s) + (W_p \cdot H_p))^{-2})}{W_p^T \cdot ((W_s \cdot H_s) + (W_p \cdot H_p))^{-1} + \lambda_p} \tag{11}$$

where $\odot$ is the element-wise product operator and the ^{T} is the transpose operator .

Once the update rules have been performed, the estimated spectrograms $\hat{X}_s$ and $\hat{X}_p$ are used to compute soft masking M_s (speech) and M_p (piano) (Wiener masking) since it provides less artifacts in the resynthesis but increases the amount of interference between speech and piano.

$$\mathbf{M}_s = \frac{\hat{\mathbf{X}}_s}{\hat{\mathbf{X}}_s + \hat{\mathbf{X}}_p} \tag{12}$$

$$\mathbf{M}_p = \frac{\hat{\mathbf{X}}_p}{\hat{\mathbf{X}}_s + \hat{\mathbf{X}}_p} \tag{13}$$

The phase information related to the speech is computed by multiplying the mask M_s with the complex spectrogram related to the mixture signal $x(t)$. The inverse transform is then applied to obtain an estimation of the speech signal $x_s(t)$. The computation of $x_p(t)$ is performed in a similar procedure taking into account M_p. In algorithmic approximation, the separation procedure is detailed in Algorithm 3.

Algorithm 3 Speech and Piano Separation

1 Compute the magnitude spectrogram X of the mixture signal.
2 Initialise H_s and H_p with random nonnegative values.
3 Initialise W_s and W_p from the training stage.
4 Update H_s using eq. (8) or eq. (10)
5 Update H_p using eq. (9) or eq. (11)
6 Repeat steps 4-5 until the algorithm converges (or the maximum number of iterations $MaxIter$ is reached).
7 Reconstruction of the estimated speech signal $x_s(t)$
8 Reconstruction of the estimated piano signal $x_p(t)$

3. EVALUATION

3.1. Test data

To evaluate the performance of the proposed method, we have created a test database D composed of 10 mixtures signals. Each mixture signal is composed of a 20 seconds duration speech and polyphonic piano excerpt. Each piano excerpt has been randomly extracted from the MAPS database [15]. Each speech excerpt has been randomly extracted from a set of 44 sentences spoken by a Spanish speaker. From these sentences, we have selected 10 excerpts of 20 seconds duration. Highlight that the set of syllables and piano notes used in the training are not the same used in the test in order to validate the results.

To evaluate different acoustic scenarios, the test database D has been mixed using -5, 0 and 5 dB of signal-to-noise ratio (see Table 1).

Table 1. Acoustic scenarios in the evaluation process.

Name	SNR(dB)
D_{-5}	-5
D_0	0
D_5	5

3.2. Experimental setup

The proposed method has been tested using different configurations of parameters: $N = (4096, 2048, 1024)$, $J = (2048, 1024)$, $maxIter = (100, 200, 300, 500)$. However, we have used $N = 4096(93ms)$, $J = 1024(23ms)$ and $maxIter = 100$ because a preliminary study showed that that configuration showed better results and lower computational cost.

Assuming the previous configuration (N, J and $maxIter$), separation results will be analyzed taking into account the type of divergence (KL or IS), the number of spectral patterns $K_s - K_p = (1-1, 3-1, 3-3, 5-1, 5-5, 10-1)$ and the sparsity parameters $\lambda_s - \lambda_p = (0-0, 0-1, 0.3-0.7, 0.5-0.5, 0.5-0.7, 0.5-1, 0.7-0.3, 0.7-0.5, 1-0, 1-0.5, 1-1)$.

3.3. Metrics

Three metrics [16] are used to measure the performance of the proposed method: source-to-distortion ratio (SDR), which reports about the overall quality of the separation process; source-to-interferences ratio (SIR), which provides a measure of the presence of piano sounds in the speech signal and vice versa; and source-to-artifacts ratio (SAR), which reports about the artifacts in the separated signal due to separation and/or resynthesis.

4. EXPERIMENTAL RESULTS

The proposed method was evaluated using all the possible combinations (type of divergence, number of spectral patterns and sparsity parameters) explained in section 3.2. Experimental results indicated that the best results were obtained using the optimal configurations shown in Table 2.

Table 2. Optimal configurations for speech-piano separation

Name	Divergence	K_s	K_p	λ_s	λ_p
C_1	IS	5	5	1	1
C_2	KL	5	5	1	0.5

The optimal configurations C_1 and C_2 use a number $K_s - K_p = 5$ speech and piano spectral patterns because this is the minimum number of patterns to model the spectral diversity exhibited by speech and piano (in a less proportion). Moreover, both configurations show the sparsity constraint active to improve the quality of the speech. As a consequence of the monophonic feature of speech, the speech sparsity parameter λ_s is higher that piano sparsity λ_p because speech is more sparse than piano instrument. As a example, a 6-seconds mixture spectrogram (Figure 2) and the output of the proposed method (Figure 3) using the configuration C_2 are shown. It can be observed how our proposal has successfully extracted the main features of the speech sounds in the estimated spectrogram.

Separation results are shown in Figure 4 in which the standard NMF ($\lambda_s = \lambda_p = 0$), the optimal configurations C_1, C_2 and the ideal case are compared. The ideal case shows the best SDR, SIR and SAR since in this case, the estimated speech is composed of the speech used in the mixing process to create the test database. It can be seen how all metrics (SDR, SIR and SAR) increase to evaluate a more ideal acoustic scenario. This fact is because our system performs better separation when the speech exhibits a higher power compared to the piano signal. In the three acoustic scenarios we can observe

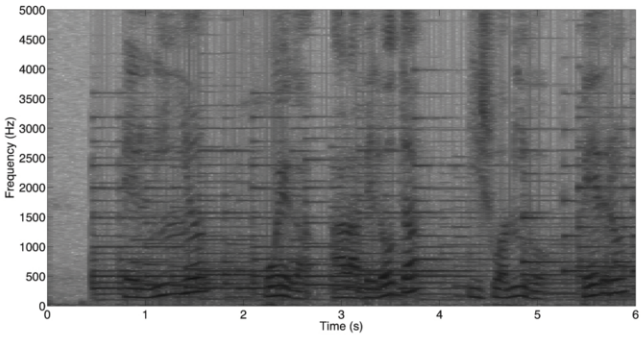

Figure 2. Time-frequency representation of a mixture composed of speech and piano sounds

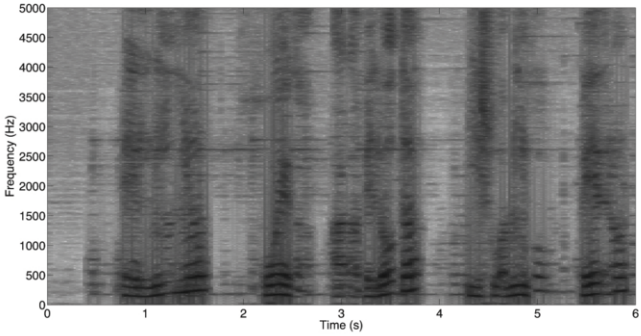

Figure 3. Time-frequency representation of the estimated speech using the configuration C_2

the configuration C_1 achieves the best SDR-SAR results considering the quality of the estimated speech but the speech contains a higher interference from piano. However, the configuration C_2 provides worse results taking into account the quality of the estimated speech (the speech is still clearly intelligible) but a lower interference from piano. Under our opinion, both configurations C_1 and C_2 can be selected as the best one because the fundamental criterion depends on the subjective quality provided by the highest SDR-SAR or SIR to each listener.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we have developed a system for separating speech from background non-stacionary noise (polyphonic piano music) in single-channel recordings. Our system, based on a supervised NMF approach, is able to learn most of spectral patterns of the syllables of the Spanish speech and the spectral patterns of the piano notes. Moreover, a sparsity constraint has been modeled to improve the separation results. An advantage of our system is its flexibility to analyze another type of non-stationary noise replacing the spectral patterns of piano by the spectral patterns of the specific non-stationary noise.

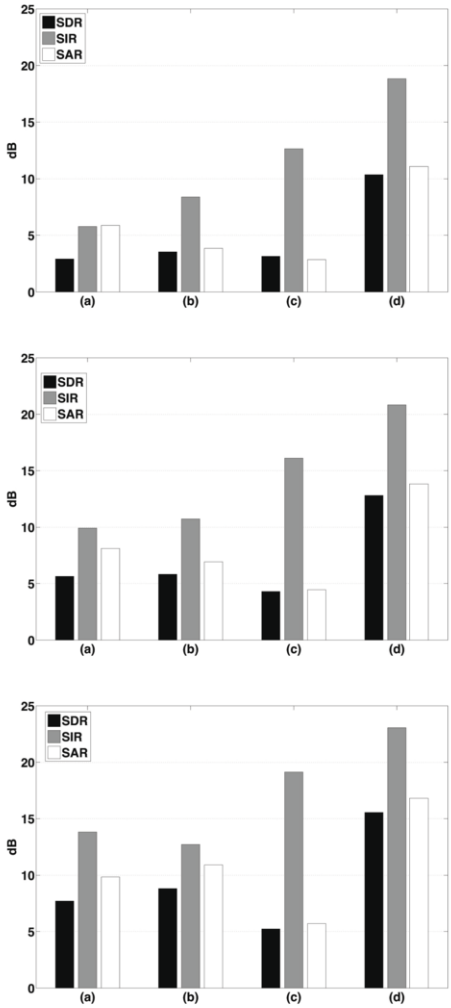

Figure 4. SDR-SIR-SAR results comparing: (a) the standard NMF ($\lambda_s = \lambda_p = 0$), (b) Configuration C_1, c) Configuration C_2 and (d) Ideal case. Test database D_{-5} (top); Test database D_0 (middle); c) Test database D_5 (bottom)

Results show that a small number of speech and piano spectral patterns is needed to model the spectral diversity exhibited by speech. The optimal configurations use the sparsity constraint to improve the quality of the speech. The configuration C_1 is the best considering the quality of the estimated speech but the configuration C_2 is the best one taking into account the minimum interference from piano.

Our future work will be focused on two topics. Firstly, developing a semi-supervised approach in order to allow the system to learn the unknown patterns active in the mixture. Secondly, a study of the influence of the speech spectral patterns in the performance of the separation taking into account different voices of different vocal characteristics.

ACKNOWLEDGEMENTS

This work was supported by the Andalusian Business, Science and Innovation Council under project P2010- TIC-6762 and (FEDER) the Spanish Ministry of Economy and Competitiveness under Project TEC2012-38142-C04-03.

References

[1] M. N. Schmidt, Jan Larsen and Fu-Tien Hsiao. Wind Noise Reduction Using Non-Negative Sparse Coding, Conference: IEEE Workshop on Machine Learning for SignalProcessing-MLSP ,2007.

[2] P. Smaragdis, B. Raj and M. Shashanka. Supervised and Semi-Supervised Separation of Sounds from Single-Channel Mixtures. Mitsubishi Electric Research Laboratories Cambridge MA, USA. Department of Cognitive and Neural Systems Boston University, Boston MA, USA. 2007.

[3] G. Mysore and P. Smaragdis. A non-negative approach to semi-supervised separation of speech from noise with the use of temporal dynamics. In Proceedings Interntational Conference on Acoustics, Speech and Signal Processing (ICASSP). Prague, Czech Republic, May, 2011

[4] M. Schmidt and R. Olsson. Single-channel speech separation using sparse non-negative matrix factorization, in *Spoken Language Proceesing, ISCA International Conference on (INTERSPEECH)*, 2006.

[5] T. Virtanen. 'Monaural Sound Source Separation by Non-Negative Matrix Factorization with Temporal Continuity and Sparseness Criteria", *IEEE Transactions on Audio, Speech, and Language Processing*, no.3, vol 15, March 2007.

[6] K. W. Wilson, B. Raj and P. Smaragdis, 2008. Regularized Non-Negative Matrix Factorization with Temporal Dependencies for Speech Denoising. In proceedings of Interspeech 2008, Brisbane, Australia, September 2008

[7] J. So-Young, K. Kyuhong, J. Jae-Hoon and C. Kwang. Semi-blind disjoint non-negative matrix factorization for extracting target source from single channel noisy mixture, *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, 2009

[8] E. Grais, H. Erdogan. Single Channel Speech Music Separation Using Nonnegative Matrix Factorization with Sliding Windows and Spectral Masks. INTERSPEECH, 2011

[9] J. Parras-Moral, J., F. Canadas-Quesada, P. Vera-Candeas and N. Ruiz-Reyes. 'Audio restoration of solo guitar excerpts using a excitation-filter instrument model, *Stockholm Music Acoustics Conference jointly with Sound And Music Computing Conference*, Stockholm, Sweden, 2013

[10] D. Lee and S. Seung. Learning the parts of objects by nonnegative matrix factorization, *Nature*, vol. 401, no 21, pp. 788-791, 1999

[11] D. Lee and H. Seung. Algorithms for Non-negative Matrix Factorization, in Ad- vances in NIPS, pp. 556-562, 2000.

[12] C. Fevotte, N. Bertin and J.-L. Durrieu. Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis. Neural Computation. 2009.

[13] http://www.zoom.co.jp

[14] Masataka Goto. Development of the RWC Music Database, Proceedings of the 18th International Congress on Acoustics (ICA 2004), pp.I-553-556, April 2004. (Invited Paper)

[15] V. Emiya, N. Bertin, B. David and R. Badeau. A piano database for multipitch estimation and automatic transcription of music. 2010.

[16] E. Vincent, R. Gribonval, and C. Fevotte. Performance measurement in blind audio source separation. IEEE Transactions on Audio, Speech and Language Processing, 14(4):1462-1469, 2006.