LETTER

On Detecting Target Acoustic Signals Based on Non-negative Matrix Factorization

Yu Gwang JIN[†a], ***Nonmember and*** **Nam Soo KIM**[†b], ***Member***

SUMMARY In this paper, we propose a novel target acoustic signal detection approach which is based on non-negative matrix factorization (NMF). Target basis vectors are trained from the target signal database through NMF, and input vectors are projected onto the subspace spanned by these target basis vectors. By analyzing the distribution of time-varying normalized projection error, the optimal threshold can be calculated to detect the target signal intervals during the entire input signal. Experimental results show that the proposed algorithm can detect the target signal successfully under various signal environments.

key words: *acoustic target signal detection, non-negative matrix factorization, normalized projection error*

1. Introduction

Target signal detection is an approach that detects the intervals where the target signal exists in the given input signal. For a number of applications in audio-based recognition and communication, it is quite often necessary to find the time intervals where the target signal of our interest resides. This becomes a challenging task if there also coexist other signals or interferences.

Voice activity detection (VAD), which decides whether or not the current input signal frame contains active speech components, is one of the most popular approaches used in speech communication systems. VAD algorithms based on statistical models have been proposed recently and demonstrated impressive performances [1]–[4]. In a wide sense, VAD can be considered a target signal detection technique where the speech is the target signal and the background noises are treated as the interferences. Even though the VAD approaches have been successful in detecting active speech signals, it is difficult to extend them to the detection of arbitrary target signals. Another possible way to detect target signals may be to apply a sequence recognition technique such as the hidden Markov model (HMM) approach widely deployed in speech recognition [5]. Since, however, this technique requires training not only the model for the target signal but also the models for the unknown interferences, it is not considered suitable for a practical implementation of general target signal detection.

In this letter, we propose a novel technique to detect acoustic target signals based on non-negative matrix factorization (NMF). NMF is a signal analysis method in which the data matrix is factorized into two constrained matrices of non-negative elements. It is quite useful for learning parts-based representation of the database by extracting a major basis which implies the data characteristics [6]. For target signal detection, the NMF basis is obtained from a training database constructed by collecting the features of the target signal. When an input signal is given, the feature extracted in each frame is projected onto the subspace spanned by the NMF basis. If the projection error is large, we can deduce that the current frame is less likely to represent the target signal. On the other hand, small projection error tells us with a high probability that the data in the current frame is actually generated from the target signal source. A robust method to determine the threshold for target signal detection is also presented. A number of experiments show that the proposed acoustic target signal detection approach performs well in various signal environments.

2. Non-negative Matrix Factorization

NMF is one of the popular techniques for multivariate data decomposition and it has been applied in broad research areas. When a collection of the input data is represented by an $n \times m$ matrix $\mathbf{V}$, it can be approximately factorized into two matrices $\mathbf{W}$ and $\mathbf{H}$ with dimensions $n \times r$ and $r \times m$, respectively, as follows [6]:

$$\mathbf{V} \approx \mathbf{WH} \qquad \text{or,}$$

$$V_{ij} \approx (\mathbf{WH})_{ij} = \sum_{k=1}^{r} W_{ik} H_{kj} \tag{1}$$

where r denotes the number of basis vectors and A_{ij} denotes the ij-th element of a matrix $\mathbf{A}$.

The way of factorizing a certain matrix is generally non-unique and a lot of methods have been developed with different constraints. NMF is different from the other methods in that it has the constraint that all the factors of $\mathbf{W}$ and $\mathbf{H}$ must be non-negative. For the factorization of an input data matrix with this constraint, we can apply the multiplicative update rules which find a suboptimal solution iteratively. Different update rules are derived depending on the distance measure defined over the space of matrices. In this work, we apply the Euclidean distance measure which results in the following update rules [7] :

Manuscript received September 24, 2009.

Manuscript revised December 21, 2009.

[†] The authors are with the School of Electrical Engineering and the Institute of New Media and Communications, Seoul National University, Seoul 151–742, Korea.

a) E-mail: ygjin@hi.snu.ac.kr

b) E-mail: nkim@snu.ac.kr

DOI: 10.1587/transinf.E93.D.922

$$H_{kj} \leftarrow H_{kj}\frac{(\mathbf{W}^T\mathbf{V})_{kj}}{(\mathbf{W}^T\mathbf{WH})_{kj}},$$
$$W_{ik} \leftarrow W_{ik}\frac{(\mathbf{VH}^T)_{ik}}{(\mathbf{WHH}^T)_{ik}}, \tag{2}$$

and in this work the initial value of each component of **H** and **W** is chosen randomly.

If we regard the r columns of **W** as the basis vectors, each column of **H** turns out to be the coefficients corresponding to the data vector projected onto the subspace spanned by these basis vectors. Since **W** is estimated from a training database of the target signal, the subspace spanned by the columns of **W** possesses representative characteristics of the target signal and it is reasonable to refer to it as the signal subspace. When a vector in the original data space is given, its projection onto the subspace constructed by the NMF basis vectors accounts for the information that can be extracted from the training data of the target signal. In contrast, the distance between the original data vector and its projection onto the signal subspace indicates how far the given data deviates from the assumed target signal source. Summarizing, we can measure the similarity between a data vector and the target signal by taking advantage of the projection error which is defined as the distance between a vector and its projection onto the signal subspace.

3. Target Signal Detection Based on NMF

In this section, we propose a novel approach to detect a specific acoustic target signal by measuring the time-varying projection error based on NMF. Our goal is to identify the time intervals during which the desired target signal exists. The proposed technique requires a training database of the target signal for the NMF analysis without a need to collect the non-target signals and interferences.

The overall block diagram of the proposed target signal detection algorithm is shown in Fig. 1. As a feature vector of the signal, we employ the magnitude spectrum which is obtained at each frame by applying the modulated complex lapped transform (MCLT) which reduces the blocking artifacts in speech enhancement [8], [9]. Even though we apply the proposed technique to MCLT for the purpose of future extension to speech enhancement, it can be successfully applied to any transform domain features such as FFT. First, we estimate the NMF basis vectors from the training database of the target signal by following (2). Once a stream of feature vectors extracted from the input signal is given, each feature vector is projected onto the signal subspace spanned by the NMF basis. By analyzing the projection error distribution obtained from the input data, we can calculate the optimal threshold for detecting the target signal. Finally, a frame-by-frame based decision as to whether the target signal exists is made by comparing the projection error with the threshold.

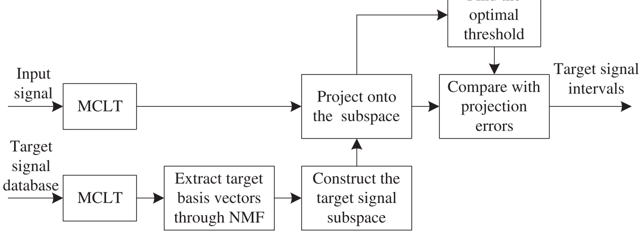
Input signal
MCLT
Project onto the subspace
Find the optimal threshold
Compare with projection errors
Target signal intervals
Target signal database
MCLT
Extract target basis vectors through NMF
Construct the target signal subspace

Fig. 1 Block diagram of proposed target signal detection system.

Let $\mathbf{x}(t)$ denote the input feature vector at time t. Then, the projection error $E(t)$ is given by

$$E(t) \equiv \min_{\alpha} \frac{\|\mathbf{x}(t) - \mathbf{W}\alpha\|^2}{\|\mathbf{x}(t)\|^2} \tag{3}$$

where $\|\cdot\|$ indicates the norm of a vector. In (3), it is noted that we normalize the conventional Euclidean distance by the input signal energy in order to make our detection algorithm immune to the signal loudness. It is well-known that the minimization on the right hand side of (3) is achieved with

$$\alpha_{opt} = \mathbf{W}^{+}\mathbf{x}(t) = (\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T\mathbf{x}(t) \tag{4}$$

where $\mathbf{W}^{+}$ denotes the pseudoinverse of **W** [10], [11]. By applying (4), we can rewrite (3) as

$$\begin{aligned} E(t) &= \frac{\|\mathbf{x}(t) - \mathbf{W}\alpha_{opt}\|^2}{\|\mathbf{x}(t)\|^2} \\ &= \frac{\|(\mathbf{I} - \mathbf{W}(\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T)\mathbf{x}(t)\|^2}{\|\mathbf{x}(t)\|^2} \end{aligned} \tag{5}$$

where **I** denotes the $n \times n$ identity matrix. From (5), we can see that the normalized projection error $E(t)$ of a vector $\mathbf{x}(t)$ is straightforwardly determined by the trained matrix **W**.

In order to find the optimal threshold for detection, we investigate the distribution of the normalized projection error based on the histogram of $E(t)$ over the input data. The histogram is obtained by first dividing the whole range of $E(t)$ into N non-overlapping subranges and then accumulating the number of frames at which $E(t)$ falls on each subrange. The number of subranges N should be carefully determined to make a good compromise between a fine resolution and robustness.

Since the histogram approximates the actual distribution of the time-varying projection errors over the input signal, its shape provides an important cue for finding the optimal threshold of detection. A typical example of the histogram is shown in Fig. 2 where the input signal is constructed by concatenating 5 different signals, and we can find several peaks which are separated with each other as shown in Fig. 2 (c). Without loss of generality, we can deduce that the samples falling on the first peak are generated from the target signal source. Therefore, the optimal threshold for detecting the target signal should be located between the first and second peaks. For a systematic way to find this separating position, we approximate the first and second peaks by individual Gaussian distributions. Let P_1 and σ_1 respectively represent the mean and standard deviation

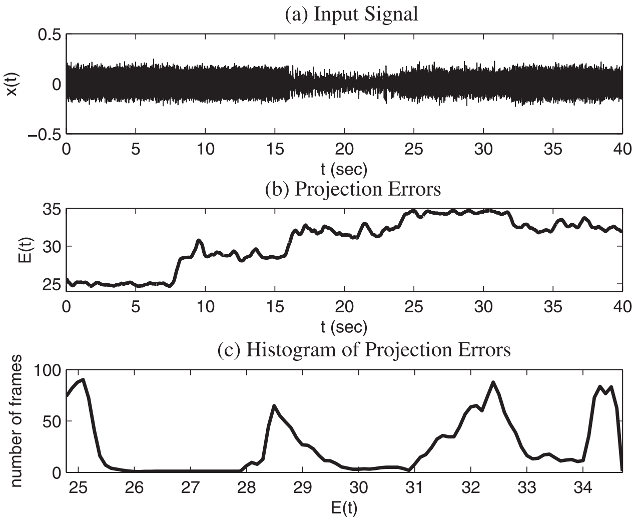

Fig. 2 An example of experimental results with a test data concatenating 5 different kinds of signals. (a) input signal, (b) projection errors, (c) histogram of projection errors ($N = 100$).

of the Gaussian distribution that approximates the first peak, and P_2 and σ_2 be those corresponding to the second peak. Then, the threshold T is determined as follows :

$$T = \beta(P_1 + \gamma_1\sigma_1) + (1 - \beta)(P_2 - \gamma_2\sigma_2) \qquad (6)$$

where β, γ_1 and γ_2 are experimentally determined constants. By varying the parameters, β, γ_1 and γ_2, we can make a different trade-off between the detection and false alarm performances. In this study, we set the parameters such that β = 0.5, and $\gamma_1 = \gamma_2 = 1.7$. Once T is determined according to (6), each feature vector $\mathbf{x}(t)$ is decided to have come from the target signal source if $E(t)$ is below T.

4. Experimental Results

In order to evaluate the performance of the proposed algorithm, we conducted a series of acoustic target signal detection tests under various signal environments. Instead of simulating the real environment, we simply concatenated or added a number of different signals and tried to detect the target signal among them. For the tests, the white, HF channel, car interior, destroyer operations room and jet cockpit noises from the NOISEX-92 database [12] were applied. For each test file, a single type of noise was treated as the target signal while the others were considered as unwanted signals or interferences. Each file was sampled at 8 kHz, and the target basis vectors for each type of target signal were trained though NMF based on a training data of length 16 seconds. MCLT was applied to obtain the magnitude spectrum at each frame whose window size was 256 samples, and the magnitude spectra of 10 adjacent frames were appended to form a feature vector to reflect time-varying input characteristics.

First, a test on target signal detection was performed with various r, the number of basis vectors. This experiment was carried out to investigate the relation between the target signal detection probability P_d and the number of basis vectors. For this test, the four signals, HF channel, car interior, destroyer operations room and jet cockpit noises were applied as the target signals and the white noise was used as the background noise. The length of each test file was 40 seconds in which five segments of the target signal, each of length 2 seconds were placed regularly. The number of basis vectors which were trained from the target signal database was varied from 1 to 30 while the signal-to-noise ratio (SNR) was maintained to 0 dB. The results are shown in Fig. 3 from which it is found that P_d of each target signal showed unstable and irregular variation when the number of basis vectors was less than 10. On the other hand, when r exceeded 10, a stable value of P_d was obtained. From these results, we can see that few basis vectors cannot separate the target signal sufficiently from the other signals. And it is also worth noting that too many basis vectors will degrade the performance since even non-target signals have less projection error as the dimension of the signal subspace increases.

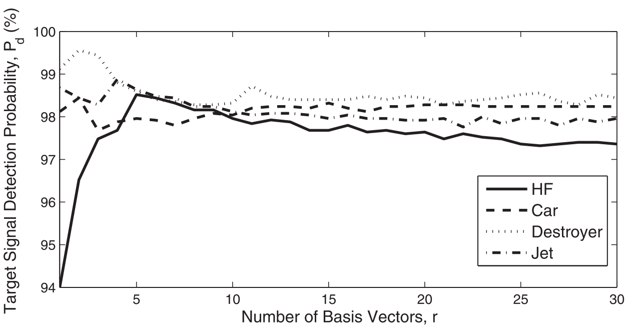

Fig. 3 Target signal detection probability (P_d) versus the number of basis vectors, when SNR = 0 dB using white noise as the background.

Table 1 Target signal detection probability (P_d), false alarm probability (P_{fa}) and probability of miss (P_m) of the proposed approach for various types of signals without additive background noises.

	P_d(%)	P_{fa}(%)	P_m(%)
white	99.84	0.00	0.80
HF	99.44	0.00	2.80
car	99.72	0.35	0.00
destroyer	99.64	0.15	1.20
jet	99.96	0.00	0.20
Avg.	99.72	0.10	1.00

Next, the target signal detection test was performed in the condition that various kinds of signals were concatenated one after another. For this test, the five different signals mentioned above were switched for every 8 seconds in a random order so that the total length of each test data became 40 seconds. The energy of each signal was kept the same so that the effect of loudness variation was eliminated. Tests were conducted for each signal, when one was selected as the target signal while the others were considered as non-target components. For each target signal, 10 basis vectors were extracted through NMF as in the previous experiment. The results are summarized in Table 1 where P_{fa} and P_m de-

Table 2 Target signal detection probability (P_d) of the proposed approach for various SNR condition using white noise as the background.

	−5 dB	0 dB	5 dB	10 dB	15 dB
HF	97.56	97.96	97.84	97.80	97.88
car	97.96	98.04	97.88	97.72	97.68
destroyer	98.68	98.28	97.84	97.96	98.12
jet	97.92	98.20	97.92	98.00	97.96
Avg.	98.03	98.12	97.87	97.87	97.91

note the false alarm probability and the probability of miss, respectively. From the result we can see that the proposed method could detect the target signal intervals almost exactly when the input target signal was not contaminated by additive background noises.

Finally, in order to verify the performances of the proposed algorithm in noisy environments, detection test was conducted by varying the SNR. In this test, 40 seconds of white noise was used as a background noise and as in the first experiment, five segments, each of length 2 seconds, of the other four noises were added by varying the SNR from −5 dB to 15 dB. Same to the previous test, 10 basis vectors were extracted for each target signal. The results are summarized in Table 2. As seen in Table 2, every type of the target signal could be detected relatively well in noisy conditions. Particularly, it is remarkable that the performance in very low SNR environment was still excellent. It is interesting to see that some of the result at lower SNR condition was better than that obtained at higher SNR though the performance difference was slight.

5. Conclusions

In this paper, we have proposed a novel approach to detect the acoustic target signal based on NMF. Target basis vectors are trained from a collection of the target signal, and the normalized projection error in each frame is calculated by projecting the input feature vector onto the subspace spanned by the target basis vectors. In order to determine the optimal threshold value, the histogram based method which approximates the distribution of the projection error is employed. Experimental results have shown that the proposed algorithm can detect the target signal successfully under various signal environments.

Acknowledgments

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology (No. 2009-0083044).

References

[1] J. Sohn, N.S. Kim, and W. Sung, “A statistical model-based voice activity detection,” IEEE Signal Process. Lett., vol.6, no.1, pp.1–3, Jan. 1999.

[2] J.W. Shin, J. -H. Chang, and N.S. Kim, “Voice activity detection based on a family of parametric distributions,” Pattern Recognit. Lett., vol.28, pp.1295–1299, Aug. 2007.

[3] J.W. Shin, H.J. Kwon, S.H. Jin, and N.S. Kim, “Voice activity detection based on conditional MAP criterion,” IEEE Signal Process. Lett., vol.15, pp.257–260, Feb. 2008.

[4] J.W. Shin, J. -H. Chang and N.S. Kim, “Voice activity detection based on statistical models and machine learning approaches” Comput. Speech Lang., vol.3, no.3, pp.205–210, March 2009.

[5] L.R. Rabiner, “A tutorial on hidden Markov models and selected applications in speech recognition,” Proc. IEEE, vol.77, no.2, pp.257–286, Feb. 1989.

[6] D.D. Lee and H.S. Seung, “Learning the parts of objects by non-negative matrix factorization,” Nature, vol.401, pp.788–791, Oct. 1999.

[7] D.D. Lee and H.S. Seung, “Algorithms for non-negative matrix factorization,” Advances in Neural Information Processing Systems, vol.13, pp.556–562, 2001.

[8] H. Malvar, “A modulated complex lapped transform and its applications to audio processing,” Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing, vol.3, pp.1421–1424, March 1999.

[9] H. Malvar, “Fast algorithm for the modulated complex lapped transform,” IEEE Signal Proces. Lett., vol.10, no.1, pp.8–10, Jan. 2003.

[10] G. Strang, Introduction to Linear Algebra, 3rd ed., Wellesley Cambridge, 2005.

[11] S.H. Friedberg, A.J. Insel, and L.E. Spence, Linear Algebra, 4th ed., Prentice Hall, 2003.

[12] A. Varga and H.J.M. Steeneken, “Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems,” Speech Commun., vol.12, no.3, pp.247–251, July 1993.