Multi-band Approach of Audio Source Discrimination with Empirical Mode Decomposition

Md. Khademul Islam Molla[1], *Keikichi Hirose*[2] *and Nobuaki Minematsu*[1]

[1]Graduate School of Frontier Sciences, [2]Graduate School of Information Science and Technology
The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan
{molla, hirose, mine}@gavo.t.u-tokyo.ac.jp

Abstract

This paper presents a content-based approach of audio source indexing without any prior knowledge about the sources. The empirical mode decomposition (*EMD*) scheme, capable of decomposing nonlinear and non-stationary signals into some bases, is employed to implement the sub-band approach of the audio discrimination technique. The feature vectors are derived from each of selected sub-bands of the target frame. Linear predictive cepstrum coefficient (*LPCC*) is used as the main feature vector and Kullback-Leibler divergence (*KLd*) is performed as the scoring function to measure the similarity of the feature vectors. The higher order statistics (*HOS*) is employed to compute the *LPCC*. The use of *HOS* makes *LPCC* less affected by Gaussian noise. The experimental results show that the sub-band approach produces better discrimination efficiency than that of the full-band technique. This discrimination method is also suitable to solve the source permutation ambiguity in separation of multiple and concurrent moving sources from the mixture(s).

1. Introduction

The emerging advancement of multimedia technology drives the need for efficient classification of the audio signals to make the content-based retrieval process more accurate and much easier from huge database. The discrimination of audio signals is a necessary task in many other potential applications: speech recognition in multi source environment, speakers' turn detection in multi-speaker situation, broadcast news indexing, audio source separation, multimedia indexing [1]. In case where only a small part of signal is known, it is possible to segregate the audio signal comes from the same source with content based audio indexing method without any prior information about the target source.

The speaker-based audio indexing is performed in [2] using second-order statistics in measuring the inter-speaker distance. The separation is performed on sequential i.e. non-overlapping audio sources and also the performances with *LPCC* and *MFCC* features are compared. They argued that with their statistical model, *MFCC* features perform better. The multi-band approach is used by [1] in audio indexing of a sequence of different audio signals. The parameters of adaptive time frequency transform are used as the features and linear discriminant analysis is performed to classify some music signals (rock, pop, jazz etc). This is a wide group of classification rather than to categorize the audio signals come from same type of sources like two male speakers. A localization based segmentation scheme of multiple concurrent speakers is proposed in [3]. They have used cepstral features of microphone array (actually using multiple arrays) signals to index the speakers based on the spatial location cue. A speakers' subspace modeling technique is introduced in [4] to index the sequential audio signals. They claimed that the speaker model construction and speaker indexing is performed simultaneously without storing the testing speech in advance.

Traditionally, the feature extraction in discrimination is performed by computing the acoustic feature vectors over the full band of the analyzing signal. In that case, the noise corruption affects all the feature components. Whereas, in multi-band approach the band limited noise does not spread to the entire feature space. In [5], it is proposed that the human auditory system processes the acoustic features from different sub-bands independently and the merging is done in some higher order processing unit to produce the final decision. The wavelet based multi-band *LPCC* features are proposed in [6] and used as the front end of speaker identification.

In this paper we have proposed a sub-band scheme of content-based audio source indexing technique in which the features of the current signal block will be used as the source model of the next one. It can efficiently be used in source-based indexing of audio signals in both sequential and concurrent situations. In the audio source separation scenario of multiple concurrent moving sources [7, 8], the proposed discrimination system is also useful to solve the permutation problem of the source signals. The empirical mode decomposition (*EMD*), a data adaptive method is employed here to implement the multi-band feature extraction scheme. The feature vector containing linear predictive cepstrum coefficient (*LPCC*) at each sub-band is computed using *HOS*, and Kullback-Leibler divergence (*KLd*) is used to score the similarity of the feature vectors. The overall score provided by the individual sub-band serves the decision factor of the discrimination. Regarding the organization of this paper, the multi-band decomposition with *EMD* method is described in section two. Section three illustrates the feature extraction and source discrimination process. The experimental results and discussion are produced in section four, and finally sections five presents some concluding remarks.

2. Multi-band Decomposition

In the proposed discrimination method original signal stream is divided into a series of blocks termed here as the analyzing signal block. Each block is represented as multiple bands using *EMD*. It is a recently developed method, specifically designed to analyze the non-linear and non-stationary properties of a time domain signals [9]. The principle of the *EMD* technique is to decompose any signal into a sum of the oscillatory components called intrinsic mode functions (*IMF*s). Each *IMF* satisfies two conditions: (i) in the whole data set the number of extrema (maxima and minima) and the number of zero crossing must be same or differ at most by one, (ii) at any point, the mean value of the envelope defined by the local maxima and the envelope defined by the local minima is zero. The first condition is similar to the narrow-band requirement for a stationary Gaussian process and the second condition is a modification of global requirement to a local one, and is necessary to ensure that the instantaneous frequency will not have redundant fluctuations as induced by asymmetric waveforms [9]. The *EMD* is also interpreted as dyadic filer-bank [10]. For any time series *s(t)* the *EMD* algorithm can be expressed as follows:

1. Initialize the residual $r_0=s$ and index of *IMF* $j=1$
2. Compute j^{th} *IMF*
3. I) set $g_0=r_{j-1}$ and $i=1$
   II) Identify the extrema (minima and maxima) of g_{i-1}
   III) Compute upper and lower envelopes h_{i-1} and l_{i-1}
   IV) Find mean envelope $m_{i-1}=(h_{i-1}+l_{i-1})/2$
   V) Update $g_i=g_{i-1}-m_{i-1}$ and $i=i+1$
   VI) Repeat steps (II)-(V) until g_i being an *IMF*. If so, the j^{th} *IMF* $f_j=g_i$
4. Update residual $r_j=r_{j-1}-f_j$
5. Repeat steps 2 to 3 with the index of *IMF* $j=j+1$

At the end of the decomposition the signal s*(t)* is represented as:

$$s(t) = \sum_{j=1}^{n} f_j + r_n \qquad (1)$$

where n is the number of *IMF*s and r_n is the final residue. Another way to explain how *EMD* works is that it extracts out the highest frequency oscillation that remains in the signal. Thus locally, each *IMF* contains lower frequency oscillations than the one extracted just before.

The *IMF*s are interpreted as the basis vectors representing the data. In this application, the *IMF* components are used in sub-band filtering. Conventionally, the filtering is carried out in frequency domain. Any frequency domain (e.g. Fourier based) filtering method applied on nonlinear and non-stationary signal (speech and many other audio signals) eliminates some of the harmonics, which will cause the deformation of the wave forms of the fundamental modes [11]. Using *IMF*, time domain multi-band decomposition approach is implemented here. Having n *IMF*s of the above described decomposition, the high pass, band pass and low pass filtered signals of *s(t)* represented by $s_{hp}(t)$, $s_{bp}(t)$ and $s_{lp}(t)$ respectively can be defined as:

$$s_{hp}(t) = \sum_{j=1}^{p} f_j \qquad (2.1) \qquad s_{bp}(t) = \sum_{j=b}^{p} f_j \qquad (2.2)$$

$$s_{lp}(t) = \sum_{j=p}^{n} f_j + r_n \qquad (2.3)$$

This filtering method is intuitive and direct, its basis is a posteriori and data adaptive, which mean it is based on the data and also derived from data. The advantage of this time-space filtering is that the resulting band passed signals preserve the full nonlinearity and non-stationary in physical space. In this experiment we have decomposed any analyzing signal block into three band passed signals and then the features are derived on each band. With the *EMD* producing n *IMF*s, the band passed signals are composed as: $b_1(t)=f_1+f_2$, $b_2(t)=f_3+f_4$, $b_3(t)=f_5+f_6\ldots\ldots+f_n+r_n$. An audio stream (speech signal) and its three-band decomposition using *EMD* is shown in Fig. 1. Regarding the decomposition, the completeness is given by the equation (1). As a check of the completeness, the speech signal is reconstructed by adding the sub-band signals, the error is of the order 10^{-16} which is negligible in practical sense.

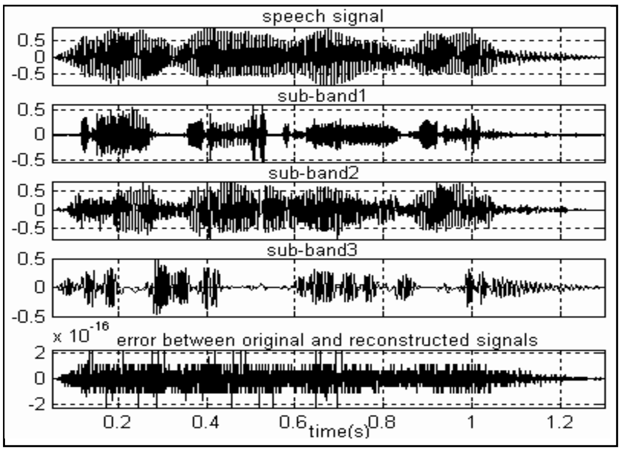

Figure 1: Three band decomposition of speech signal

3. Feature Extraction and Discrimination

The acoustic feature vector of the audio signal is computed on each of the three sub-bands. Linear predictive cepstrum coefficient (*LPCC*) is used as the principal feature in this discrimination scheme. The source indexing is performed by proper clustering of source characteristics. The schematic diagram of the proposed source discrimination method is shown in Fig 2.

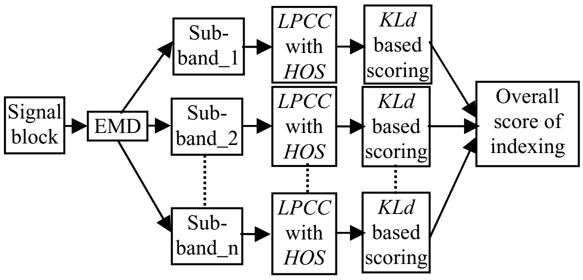

Figure 2: Schematic diagram of the proposed source discrimination model.

3.1. Computing Sub-band *LPCC*

The linear predictive cepstrum coefficient (*LPCC*) is a popular feature to capture the characteristics of audio sources. Many researches have been done by using *LPCC* as the features in speaker recognition and audio indexing [2, 6, 12]. In this discrimination method, the *LPCC* vector is computed using higher order statistics (*HOS*). Conventional *LPC* calculation is performed by solving the equation $R\alpha = \beta$ for the *LPC* coefficients α, where the elements of R represents the autoregressive (*AR*) coefficients and β is same as in equation (3). Using *HOS* the *AR* coefficients *c(i)* are reconstructed by fourth order cumulants and the *LPC*s are computed by the following method [12]:

$$C\alpha = \beta \qquad (3)$$

where $C = [c(i-j)]_{(p+1)\times(p+1)}, \beta = (E_p, 0, \cdots, 0)^T$

$$c(i) = \sum_{\tau=-M}^{M} \hat{C}_{4s}(0,\tau,\tau+i), E_p = c(0) - \sum_{i=1}^{p} a_i c(i)$$

where p is the order of *LPC* and $\widehat{C}_{4s}(\tau_1,\tau_2,\tau_3)$ is the consistent estimation of the fourth order cumulant. After computing the linear coefficients ($a_1, a_2,\ldots\ldots,a_p$) with *HOS*, the cepstral coefficients (*LPCC*) are computed in the same way as the conventional approach. The first 16 *LPCC* are used to represent the source features at any windowed time frame of a sub-band signal.

The computation of *LPC* using second order statistics like AR coefficients are sensitive to noise. Many measurement noises are with Gaussian distribution and completely characterized by its mean and variance. Consequently the higher order (above third order) cumulants of Gaussian noise are zero i.e. higher order cumulants are insensitive to Gaussian noise. This is the theoretical evidence of using *HOS* to compute *LPC*. From higher order cumulants, we can easily reconstruct the second order statistics which is used in calculating the *LPC*s.

To eliminate the channel bias, the cepstral vector is normalized by subtracting the global mean cepstral vector from each cepstral vector of individual time frame. Thus the short-term mean of the *LPCC* vectors are normalized to zero as: $\hat{x}_m(t_f)=x_m(t_f)-\mu_m$, where $x_m(t_f)$ is the m^{th} component of the *LPCC* vector at time frame t_f and μ_m is the mean of the m^{th} component of the *LPCC* vectors of a specific signal block.

3.2. *KLd* based Scoring and Discrimination

The Kullback-Leibler divergence (*KLd*) is used here as the scoring function to measure the similarity between two feature vectors in audio source discrimination technique. The *KLd*, an information theoretic distance measures the relative dissimilarities between two data distribution profiles [13]. The symmetric *KLd* between two probability mass functions $q_1(x)$ and $q_2(x)$ over a random variable is defined as:

$$KLd(q_1,q_2)=\frac{1}{2}\left(\sum_{x\in X}q_1(x)\log\frac{q_1(x)}{q_2(x)}+\sum_{x\in X}q_2(x)\log\frac{q_2(x)}{q_1(x)}\right) \qquad (4)$$

The *KLd* always takes a non-negative value, it is zero if $q_1=q_2$ and the lower value of *KLd* means high degree of similarity. The *LPCC* vectors over the entire analyzing signal block of a sub-band are considered as a single feature vector. The normalized feature vector is scaled to fit values within *-1* to *1*.

The probability mass function (pmf) of a feature vector is computed from the corresponding histogram profiles. The pmf represents the fractional contribution of the quantized feature values to the entire signal block at a specific sub-band. The result is an array with values of the feature vector falling in the interval [0, 1] which is suitable to measure the *KLd* between the model vector (feature vector of the source specific previous signal block) and the current feature vector. It is noted that the final sub-band feature is the probability mass function of selected *LPCC* components of that band over the signal block.

The model vector representing the source characteristics is derived from the given data (not priori). It is simply the feature vector derived at the starting phase of the discrimination on each sub-band as described above. Then the overall discrimination score of the k^{th} source $\eta(k)$ can be defined as:

$$\eta(k)=\frac{1}{N_b}\sum_{i=1}^{N_b}\kappa(v_i,\hat{v}_{ik}) \qquad (5)$$

where, N_b is the number of sub-bands, v_i is the feature vector of i^{th} band of the current signal block, $\hat{v}_{ik}$ represents the model feature of the k^{th} source at i^{th} band, and $\kappa(.)$ measures the *KLd* between those feature vectors. The k with minimum $\eta(k)$ is selected as the candidate source of the current signal block. To discriminate the sequential alternate sources, it is necessary to settle a threshold of $\eta(k)$ scoring efficiently to determine the region of source transition. The threshold is determined by comparing scoring of various source signals. When the model vectors of all the sources are already computed, the indexing method is similar to the concurrent source model as explained in the following section.

4. Experimental Results and Discussion

The proposed audio source indexing technique is evaluated to discriminate the audio streams of two male speech signals (ml1 and ml2), one female speech (fm) and a musical instrument (flute, fl). All signals are sampled at 16 kHz with 16 bits amplitude resolution. A section of the input stream with length of 600ms is taken as the analyzing signal block. To construct the model features, the selected signal block is preprocessed to detect the silence frames. The short term (frame of 10ms length and 5ms overlapping) energy is compared with a predefined threshold (-30dB) for this task. If more than 30% fames of a selected block are silence, that block is disregarded as it is inefficient to represent the source characteristics any more. Otherwise, the model features (vectors) are updated. We have addressed here the two scenario of source indexing as given below.

4.1. Indexing of Sequential Sources

The model features are derived from analyzing block if it contains reasonable signal energy. Then the current block is verified using equation (5) whether it belongs to the same source just previously selected. If so, the indexing continues and the model features are updated. If not, the block is verified that it belongs to one of the other previously indexed sources, and the block is grouped to the matched one. Otherwise, the $\eta(k)$ with all the indexed sources exceeds the threshold limits, a new source model is constructed from the signal block. The indexing performance of sequential sources as a form of confusion matrix (each entry represents the number of analyzing blocks) is shown in Table 1 together with the discrimination accuracies.

Table 1: The indexing performance of sequential sources as a form of confusion matrix and discrimination accuracy for both multi-band (DAm) and full-band (DAf) approaches.

	ml1	ml2	fm	fl	DAm(%)	DAf(%)
ml1	36	2	0	0	94.73	89.74
ml2	2	37	1	0	92.50	85.00
fm	0	1	34	0	97.14	94.28
fl	0	0	0	32	100.0	90.62
Overall performance					96.09	89.91

4.2. Indexing of Concurrent Sources

The indexing of concurrent sources can be applicable to resolve the permutation ambiguity in moving source separation. Suppose there are four audio sources moving in the azimuth plane (not overlapped at the same azimuth angle) subject to be separated from two mixtures. At any given time the separation is performed based on the source locations [7]. Thus four location based vector containing different source blocks are produced. The situation is represented by Figure 3.

For any given time slot, the signal blocks of different location vectors come from individual source. After computing the model features of each source, the blocks of a time slot are sorted as: any block is compared with individual source by using equation (5) and the source with minimum score is selected as the belonging one. The next block is compared with the rest of the sources and this process is continued until all the blocks are sorted. The model features are updated if the current block contains a reasonable amount of energy. The performance of concurrent source indexing is shown in Table 2.

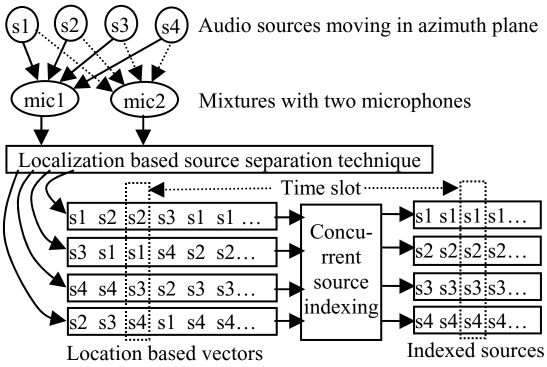

Figure 3: Concurrent source indexing situation

Table 2: The indexing performance of concurrent sources as a form of confusion matrix and discrimination accuracy for both multi-band (DAm) and full-band (DAf) approaches.

	ml1	ml2	fm	fl	DAm(%)	DAf(%)
ml1	36	1	1	0	94.73	89.74
ml2	2	38	0	0	95.50	87.70
fm	1	0	34	0	97.14	91.42
fl	0	0	0	32	100.0	96.87
Overall performance					96.84	91.43

In the two above mentioned applications we have compared the indexing performance of multi-band and full-band approaches. It is observed that multi-band approach performs better. It is also noticed that more misclassifications are occurred between two male speech signals (ml1 and ml2). Being same type of sources they may produce a closer shape of probability mass functions of the *LPCC* features. Another factor of indexing performance is the size of the analyzing signal block. The lager the block size produces more efficient representation of source characteristics. On the other hand, the smaller block size can efficiently detect the source transition point. There is a tradeoff when choosing the size. Considering both evidence we have selected it with length 600ms. It is assumed that no source changing is occurred within the chosen time slot. The proposed data adaptive filtering approach and the use of *HOS* to construct the source features make the audio source discrimination technique more noise robust and efficient in application domains.

5. Conclusions

The proposed method is able to discriminate the audio sources without any prior knowledge about the source models. The *EMD*, a data adaptive signal decomposition technique, is employed here to implement the sub-band approach. Contrary to the other harmonic analysis (e.g. Fourier, wavelet), it is well fitted to decompose nonlinear, non-stationary and non-harmonic signals. The experimental results of the proposed method are noticeable and can effectively be used in audio indexing in multimedia applications. Also the performance is compared with the full band approach and it is observed that multi-band scheme is more effective in distinguishing the audio sources. To examine the effects of the number of sub-bands, the compare the performance of other features like *MFCC* and to perform more experiments with the signals recorded in real-world noisy environments are the future targets of this research.

6. References

[1] Umapathy, K., Krishnan, S. and Jimaa, S., ``Multigroup Classification of Audio Signals using Time-Frequency Parameters``, *IEEE Trans. on Multimedia*, Vol. 7, No. 2, pp: 308-315, April 2005.

[2] Delacourt, P. and Wellekens, C., ``Audio data indexing: use of second-order statistics for speaker-based segmentation``, *Proc. of IEEE Int. Conf. on Multimedia Computing and Systems*, Vol. 2, pp: 959-963, 1999.

[3] Lathoud, G., McCowan, I. A. and Moore, D. C., ``Segmenting Multiple Concurrent Speakers using Microphone Arrays``, *Proc. of EUROSPEECH 2003*, pp: 2889-2892, 2003.

[4] Nishida, M. and Ariki, Y., ``Speaker Indexing for News Articles, Debates and Drama in Broadcasted TV Programs``, *Proc. of IEEE Int. Conf. on Multimedia Computing and Systems*, Vol. 2, pp: 466-471, 1999.

[5] Allen, J. B., ``How do humans process and recognize speech?``, *IEEE Transaction on Speech and Audio*, 2(4), pp: 567-577, 1994.

[6] Hsieh, C. T., Lai, E and Wang, Y. C., ``Robust Speech Features based on Wavelet Transform with Application to Speaker Identification``, *IEE Proceedings – Vision, Image and Signal Processing*, 149(2), pp: 108-114, 2002.

[7] Molla, M. K. I., Hirose, K. and Minematsu, N., `` Audio Source Separation by Source Localization with Hilbert Spectrum``, To appear in the *Proc. of IEEE Int. Symposium on Circuits and Systems (ISCAS'05)*, May 2005.

[8] Rickard, S. and Yilmaz O., ``On the approximate W-disjoint orthogonality of speech``, *Proc. of ICASSP'02*, 2002.

[9] Huang, N. E. et al., ``The Empirical mode decomposition and Hilbert spectrum for nonlinear and non-stationary time series analysis``, *Proc. Roy. Soc. Lond. A*, Vol. 454, pp: 903-995, 1998.

[10] Flandrin, P., Rilling, G. and Goncalves, P., ``Emperical Mode Decomposition as a filter bank`` *IEEE Sig. Proc. Letter*, Vol. 11, No. 2, pp: 112-114, Feb 2004.

[11] Huang, N. E., Wu, M. L., Qu, W., Long, S. R. and Shen, S. S. P., ``Application of Hilbert-Huang transform to non-stationary financial time series analysis``, *Applied Stochastic Model in Business and Industry*, Vol. 19, pp: 245-268, 2003.

[12] Ma, J., and Gao, W., ``Robust Speaker Recognition Based on High Order Cumulant``, *Proc. of International Conference on Spoken Language Processing (ICSLP2000)*, pp. 278-285, 2000.

[13] Kasturi, J., Acharya, R. and Ramanathan, M., ``An information theoretic approach for analyzing temporal patterns of gene expression``, *Bioinformatics*, Vol. 19, no. 4, pp: 449-458, 2003.