INTERNATIONAL JOURNAL OF APPLIED ENGINEERING RESEARCH, DINDIGUL
Volume 2 No 2 2011

REVIEW ARTICLE ISSN - 0976-4259

A new codebook design schema for VQ-based Monaural Speech-Music Segregation

Seyed-Hossein Alavinia, Farbod Razzazi
Department of Electrical Engineering, Science and Research Branch,
Islamic Azad University, Tehran, Iran,
h_alavinia@yahoo.com

ABSTRACT

Several ideas have been introduced to improve monaural speech-music segregation problem. Schema-driven approaches employ some statistical methods to model the underlying source signals. Although schema-based techniques present a high quality segregated speech and music outputs, the computational complexity is the main drawback of these methods. In this paper, we proposed an optimized version of hybrid PCA-VQ model based on K-means clustering to overcome this deficiency. K-means algorithm does not work well for high dimensional data in terms of computational complexity and curse of dimensionality issues. To overcome the computational complexity of K-means algorithm and obtain uncorrelated and the most descriptive variables, we used Principal Component Analysis (PCA) technique. This technique is a commonly used statistical approach for dimension reduction. The goal of PCA is a linear mapping which maps data to a lower dimensional space, so that variance of the data in new space is maximized. First, we employed PCA on 2-D framed STFT of the input signal to compute uncorrelated feature vectors efficiently and subsequently. Then we introduced and evaluate a modified version of k-means method for clustering. The simulation results demonstrate that the proposed schema can improve the segregated procedure in terms of PESQ criterion and complexity measures.

Keywords: K-means clustering, monaural segregation, PESQ, principal component analysis, schema-driven, vector quantization.

1. Introduction

The research on speech segregation has become an increasingly popular topic in the field of signal processing. Specifically, monaural speech-music segregation has attracted exclusive attention as a main case of the speech enhancement problem in recent decades. This technique can be divided into two categories: first, primitive date-driven methods which are also known as source-driven methods and second, knowledge-based schema-driven methods also called as model-driven methods. Primitive date-driven approaches aims to extract the underlying sources signals from mixed signal without any prior knowledge of the speakers; while in the schema-driven techniques, separation is completely based on prior knowledge of speakers.

The first class is categorized into two main methods named computational auditory scene analysis (CASA) approaches and blind source separation (BSS) methods. In CASA approaches, the mixed signal is transformed by an appropriate transformation (such as short time Fourier transform (STFT)) and is segmented into cells. Then, the cells that are belonged to one source are grouped based on some related psychoacoustic clues (such as onset, offset, and pitch period) into separated streams (Hu & Wang, 2004). CASA methods have several major drawbacks as follows: First, identifying the priority rules, which are related to grouping cues is sometimes hard. Second, they can hardly mitigate crosstalk and at last, they can hardly segregate unvoiced regions (Hu & Wang, 2004).

The main condition to apply BSS is that the number of microphones must be larger or equal to the number of sources and BSS methods considerable deteriorate when they are applied in the case of one microphone audio separation (Bermond & Cardoso, 1999).

In model-driven techniques, the original spectra of each speaker are replaced by a statistical model, which is estimated during the training phase. Then, an approximation to the original spectra is found by decoding the states, components, or code-vectors that satisfy a minimum distortion criterion. In these methods, the underlying source signals are modeled using well-known statistical speech modeling such as Gaussian Mixture Models(GMM) (Reddy & Raj, 2004), Gaussian Scaled Mixture Model (GSMM) (Benaroya & Bimbot & Gribonval, 2006), Hidden Markov Models (HMM) (Roweis, 2000), or VQ (Rowies, 2003). Among these approaches, GMM has been introduced as favorable choice in speech-music separation applications. GMM model may result in some interference signals which degrade the perceptual quality of separated signals in addition to result some over-estimation error. In addition, GMM method significantly fails, when the phase information is employed for separation and results in bad separated output signals (Asgari & Fallah & Mehrizi & Mostafavi, 2009). Presenting a compact model for phase value is usually a difficult task with unsatisfactory results (Radfar & Dansereau, & Sayadiyan, 2007). In (David, 2002) a comparison between VQ and GMM is presented in terms of optimum number of components, the number of iterations using EM (Expectation-Maximization) algorithm and overall system performance. It has been observed that the system based on VQ presents better results.

As it can be deduced from above comparisons, VQ method presents a better efficiency and high quality signal segregation in comparison with other mentioned approaches. In VQ method, the input is divided into several vectors and then each vector is mapped to the code-words of a codebook (Lu & Chang, 2010). One of the most popular and efficient clustering methods is the k-means method (Lloyd, 1957) which use prototypes (centroids) to represent clusters by optimizing the squared error function (Wallace, 1989).

In this paper, we used squared Euclidean distance criterion for clustering, which each center is the mean of the some points in that cluster. K-means algorithm does not work well for high dimensional data in terms of computational complexity and curse of dimensionality issues (Ding & He, 2004). To overcome the computational complexity of K-means algorithm and obtain uncorrelated and the most descriptive variables, several methods have been proposed over the years. Undoubtedly, the most popular approach is PCA, introduced by Karl Pearson in 1901 (Napoleon & Pavalakodi, 2011).

We applied PCA technique which is a data analysis technique that transforms a number of possibly correlated variables into a smaller number of uncorrelated variables which are called Principal Components (PCs). In PCA procedure, PCs are calculated using the eigen-value decomposition of a data covariance matrix. This technique is a commonly used statistical technique for dimension reduction. The goal of PCA is a linear mapping which maps data to a lower dimensional space, so that variance of the data in new space is maximized (Radfar &

Dansereau & Sayadiyan, 2007). After PCA analysis, K-means algorithm may be applied into the data in the new space to cluster the samples. In this paper, we propose a modified version of K-means clustering to establish source models.

The remainder of this paper is structured as follows: in the following section the existing methods for schema-based separation are briefly discussed. Section 3, introduces our proposed method. The experimental setup and simulation results are described in Section 4. Finally, Section 5 concludes the paper.

2. Schema-based Separation

In order to extract the underlying source signals from a mixed signal which consists of sum of both signals; one must find the estimations for both signals from the mixed signal. Let a mixed signal be consists of speech and music signals as follows:

$$x(n)=m(n)+s(n) \quad n=1,...,N \tag{1}$$

where *m(n)*, *s(n)* and *x(n)* are music signal, speech and the obtained mixture signal, respectively. Taking STFT from both sides of (1) and due to linearity of this transform we arrive at:

$$X(n,f)=M(n,f)+S(n,f) \tag{2}$$

where *n* and *f* are the frame number and the frequency index in a time-frequency representation, respectively. As mentioned before, in schema-driven techniques, the original spectra of each speaker is replaced by a statistical model, which is estimated during the training phase. Then, an approximation to the original spectra is found by decoding the states, components, or code-vectors that satisfy a minimum distortion criterion. Among these approaches, GMM has been introduced as favorable choice in speech-music separation applications. The covariance matrices $\sum_m$ and $\sum_s$ are assumed to be diagonal, with running element $\sigma_m^2(f)$ and $\sigma_s^2(f)$ respectively (Benaroya & Bimbot & Gribonval, 2006). In mixture mode in (Wallace, 1989). Bayesian estimation of M(n,f) and S(n,f) will be as follows:

$$\hat{M}(n,f)=\frac{\sigma_m^2(f)}{\sigma_m^2(f)+\sigma_s^2(f)}X(n,f) \tag{3}$$

$$\hat{S}(n,f)=\frac{\sigma_s^2(f)}{\sigma_m^2(f)+\sigma_s^2(f)}X(n,f) \tag{4}$$

Such GMM model may result in some interference signals which degrade the perceptual quality of separated signals in addition to result some over-estimation error. In addition, GMM method significantly fails, when the phase information is employed for separation and results in bad separated output signals in this case (Asgari & Fallah & Mehrizi & Mostafavi, 2009). Presenting a compact model for phase value is usually a difficult task with unsatisfactory results (Radfar & Dansereau & Sayadiyan, 2007). In the proposed VQ-based approach in (Asgari & Fallah & Mehrizi & Mostafavi, 2009), some phase vectors are inserted at each VQ entries in order to separate high quality signals. These vectors are used only in the synthesis stage. In addition, GMM computational complexity is high in practice. In (Lotia & Khan, 2011), a VQ-based GMM model has been proposed in order to reduce the computational complexity by reducing the parameters of GMM. In (Li & Guan & Wang & Xu & Liu, 2010) and (Radfar & Dansereau & Sayadiyan, 2007), two masks based on the estimated VQ have been produced by the MAX-VQ separation system. These estimated masks often provide corrupted re-synthesis signals with undesirable crosstalk effects.

As it can be deduced from above comparisons, VQ method presents a better efficiency and high quality signal segregation in comparison with other mentioned approaches. In VQ method, the input is divided into several vectors and then each vector is mapped to the code-words of a codebook (Lu & Chang, 2010*)*. K-means clustering is one of the most famous VQ algorithms. In this paper, we used squared Euclidean distance criterion for clustering, which each center is the mean of the points in that cluster. To overcome the computational complexity of K-means algorithm for high dimensional data, we applied PCA technique which is a data analysis technique that transforms a number of possibly correlated variables into a smaller number of uncorrelated variables which are called Principal Components (PCs). In PCA procedure, PCs are calculated using the eigen-value decomposition of a data covariance matrix. This technique is a commonly used statistical technique for dimension reduction. The goal of PCA is a linear mapping which maps data to a lower dimensional space, so that variance of the data in new space is maximized (Radfar & Dansereau & Sayadiyan, 2007). Due to linearity of this method, the equality (2) will be remained unchanged in the new domain.

3. Proposed Method

After obtaining the absolute value of STFT of both speech and music input signals, framing can be performed on the resulted matrix. The overlap value is usually half of the length of each frame. Later on, we employed PCA technique on the result.

After framing the output of absolute value of STFT function, in the first stage of PCA technique, we took the mean of the obtained function as:

$$\theta = \left(\frac{1}{m} \sum_{i=1}^{m} (\mathrm{framedS}_x) \right) \qquad (5)$$

where θ and framedS_x denote the mean vector and framed STFT of input matrix, respectively. After that, we subtracted the θ vector from each column of framedS_x matrix. We call the resulting matrix as ϕ. Then, we calculated the eigen-values and eigenvectors of the covariance matrix of $framedS_x$ matrix. For 2D data, the covariance matrix has two dimensions, since STFT matrix is a 2D operator, the covariance matrix will be a 2×2 matrix. For this matrix, the eigen-values and eigenvectors can be calculated, because the covariance matrix is a square matrix. The eigen-values are indicatives of components variances.

After that, we should sort the eigenvectors in terms of more descriptive eigenvectors and form a matrix that contains features vectors. We call this matrix as the principal component coefficients (PCs) matrix as follows:

$$PCs=(EV_1\ EV_2\ EV_3\ ....\ EV_n) \qquad (6)$$

Where EV_i denotes the i^{th} eigenvector. Then, the obtained new data in PCA space can be expressed as:

$$\psi = \phi \times PCs \qquad (7)$$

which the transformed data in PCA space denoted with ψ. In this paper we take the speech feature matrix to transform speech, music and mixture signals into PCA space. PCA is a linear operator. As mentioned before, by using PCA analysis, high dimensional data are transformed into lower dimensional data. After finding the principal components, the reduced data set is clustered by applying k-means clustering. Clustering is grouping samples based on their similarity, while samples in different groups are dissimilar (Napoleon & Pavalakodi, 2011). VQ is an efficient method for data clustering which has been widely used in different applications. In VQ method, the input is divided into several vectors and then each vector is mapped to the code-words of a codebook (Napoleon & Pavalakodi, 2011). One of the most popular VQ algorithms is K-means clustering. In this paper, we use Squared Euclidean distance criterion for clustering which each centroid is the mean of the points in that cluster. As said before, this common k-means algorithm takes much time and doesn't work well for high dimension data. In order to obviate these major drawbacks, we aim at optimizing the K-means algorithm.

3.1 Modified PCA-Kmeans Algorithm

As mentioned in former section, the PCs matrix is a square matrix which each column of it contains the coefficients of one principal component, the columns are in order of decreasing components variances. We can take some of columns of PCs matrix instead of the whole matrix. By using the eigen-values of the covariance matrix of $framedS_X$ in PCA analysis, we can determine how many column of PCs matrix should be selected.

The Main idea of the Modified PCA-Kmeans algorithm is storing the whole PCA transformed vector in the codebook, while recruiting clusters for the new vectors by the most descriptive eigen-vectors.

Let the vector, which contains the the eigen-values of covariance matrix of $framedS_X$ in PCA analysis be α, and the number of selected columns of PCs matrix be C, if the percent value of the energy of the desired information be denoted as P, the relationship between P and α will be:

$$P=\frac{sum(\alpha(1:C))}{sum(\alpha)} \qquad (8)$$

We can take the C-first column of PCs matrix instead of the whole matrix and apply it in k-means algorithm to decrease the dimension of data. However, we proposed another method which gives us better results in comparison to this method. In general, we can summarize K-means algorithm into two stages (1) The clusters recruitment stage and (2) re-estimating the

REVIEW ARTICLE ISSN - 0976-4259

centroid of each cluster. In order to improve the quality of separation, we performed the recruitment stage by using the C-first column of PCs matrix, and applied the whole PCs matrix in re-estimating the centroid of each cluster stage. We call this proposed method as Reduced Clusters Recruitment, Complete Re-estimation of Centroids (RC3) training procedure.

In PCA-Kmeans method, we employed C-first column of PCs matrix in CBs generation algorithm to transform the underlying signals into PCA space. Therefore, each CBs has K rows and C columns, whick K is the number of clusters in k-means algorithm. Likewise, in separation algorithm, mixed signal is transformed into PCA space by C-first columns of PCs matrix and then, these transformed vectors are compared with code-words to select the appropriate code-words.

In RC3 method, the whole PCs matrix is employed to generate both of CBs. Thus, each CB has K rows and the number of columns of each CB is equal to the number of columns of PCs matrix. In addition to, in separation algorithm, we transformed the mixed signal into PCA space similar to PCA-Kmeans algorithm, but this transformed matrix was compared with the C-first columns of each CB. After selection the appropriate indices of code-word, we selected the complete code-words corresponded with these indices of CBs to generate the separated speech and music signals in output.

RC3 method results in larger code-words in both speech and music codebooks. This is because of employing all entries for each transformed vector. This is due to the fact that the dimension of the codebook is larger and we replaced a more accurate codeword to estimate the separated speech and music. Therefore, the quality of separated signals will be higher than PCA-Kmeans method. However, the smaller size of vectors in comparisons will make the system less complex and with less memory usage.

3.2 Segregation Algorithm

As it is observed in Eq. (2), the STFT of the mixture signal is equal to the sum of STFTs of the underlying sources signals which the mixture signal depends on both amplitude and phase of the STFT of speech and music signals. In (Asgari & Fallah & Mehrizi & Mostafavi, 2009), by using Maximum likelihood Amplitude Estimator (MLAE), the optimum approximation for DFT amplitude is derived as the sum of the speech and music spectrum amplitudes. Generalizing this result to STFT, Eq. (1) can be expressed as:

$$|X(n,f)| \approx |S(n,f)| + |M(n,f)| \qquad (9)$$

In the segregation process, after framing the mixture's STFT amplitude and transforming the resulted mixture to PCA space, the algorithm should find the closest approximation of the mixture signal among the sum of both speech and music code-words of the obtained codebooks. The search algorithm in codebooks can be formulated as bellow:

$$p,q = \underset{p^*,\, q^*}{\operatorname{argmin}} \left\{ \left\| \left(|\psi_{mix}| - (|CBs_{p^*}| + |CBm_{q^*}|) \right) \right\|_1 \right\} \qquad (10)$$

where p and q are the code-words indices in each codebook and ψ_{mix} denotes the transformed mixture to PCA space. After codebookwhere p and q are the code-words indices in each codebook and ψ_{mix} denotes the transformed mixture to PCA space. After the appropriate codeword selection stage, the previous stages of the selection stage should be reversed in order to extract the separated signals. Finally, to calculate the original data from obtained matrices in this procedure, we have:

$$\beta=(PCs\times\psi')+(framedS_x-\phi) \qquad (11)$$

where β is the reconstructed data. In figure 1, the whole procedure of segregation in case of STFT for PCA-Kmeans and RC3 methods is illustrated.

4. Experimental Results

In the paper, we assess the quality of gained results in terms of The Perceptual Evaluation of Speech Quality (PESQ) criterion, as an objective speech quality model, which was approved as ITU-T recommendation P.862 (ITU-T Rec P.862, 2001). This measure is highly correlated with the Mean Opinion Score (MOS) as one important metric, but evaluation of MOS is very expensive and time-consuming for subjective determination of quality (Cernak & Rusko, 2005). The range of the PESQ score is -0.5 (which indicate the several degradation) up to 4.5 (which means that the measured speech has no distortion and is exactly the same as the original signal).

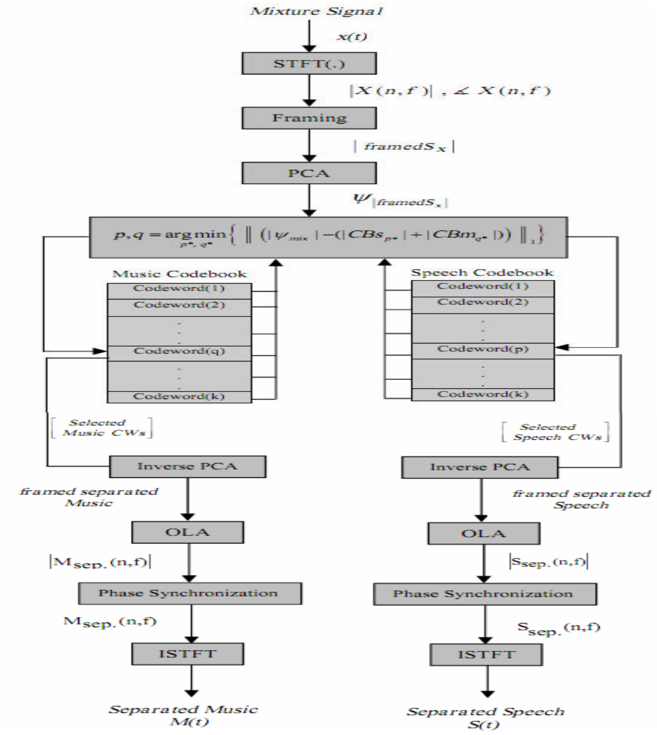

Figure1: The speech-music segregation procedure in STFT for PCA-Kmeans and RC3 clustering.

We have compared the ordinary PCA-kmeans clustering with the RC3 clustering method, as the proposed method, to extract speech and music signals from mixture signal in both FFT and STFT domains. These four cases are evaluated in terms of the search time into codebooks as the computational complexity measure and PESQ of the speech signal as the quality measure. Two codebooks were constructed for the speech signals of a male speaker and the music signals including piano segments which the sizes of both codebooks are M=512 clusters.

For both of the operations, the codebook generation and speech/music signals separation, Eq. (5) was used to calculate the vector θ associated with speech STFT and Eq. (8) was employed to obtain the parameter C for the PCs matrix related to speech signals as the feature vectors to satisfy P=0.95. Table 1 describes the PESQ rate and time consumption as computational complexity criterion for both methods: PCA-Kmeans and RC3 in FFT and STFT domains. As it is observed in this table, RC3 improved PESQ value and decreased the time of search in codebook for both domains in comparison to PCA-Kmeans method. Likewise, these methods indicate the better performance in STFT in comparison with FFT domain.

Table 1: PESQ rate and codebook search time for proposed method in FFT and STFT domains.

Domain	Method	PESQ	CB Search Time
FFT	PCA-kmeans	2.51	11 min
	RC3	2.63	19 min
STFT	PCA-kmeans	2.91	5 min
	RC3	3.02	5 min

5. Conclusion

In this paper, a novel method was proposed to cope with performance and computational complexity drawbacks of STFT schema-based monaural speech-music segregation technique. To evaluate the efficiency, we employed our method in both codebook generation segregation stages into FFT and STFT domains. This procedure generated more accurate codebooks than using standard PCA and K-means stages sequentially. The experimental results showed a slight improvement on output speech quality in addition to considerable reduction in computational complexity.

6. References

1. G. N. Hu and D. L. Wang (2004), “Monaural speech segregation based on pitch tracking and amplitude modulation,” IEEE Trans. Neural Netw, 15(5), pp 1135–1150.

2. O.Bermond and J.-F. Cardoso (1999), “Approximate likelihood for noisy mixtures,” in Proc. ICA’99, Aussois, France, pp 325-330.

3. Reddy, A.M., Raj, B (2004), "A minimum mean squared error estimator for single channel speaker separation". In: INTERSPEECH-2004, pp 2445–2448.

4. Laurent Benaroya, Frédéric Bimbot, and Rémi Gribonval (January 2006), "Audio Source Separation With a Single Sensor" IEEE Transaction on Audio, Speech, and Language Processing, 14(1)

5. Roweis, S. T (2000), "One microphone source separation. In: Proc. Neural Information Processing Systems", pp 793–799.

6. Rowies, S. T (2003), factorial models and refiltering for speech separation and denoising. In: EUROSPEECH-03, May 2003, 7, pp 1009–1012.

7. M. Asgari, M. Fallah, E. Abouie Mehrizi, A. Mostafavi (2009), "A VQ-based single-channel audio separation for music/speech mixture", UKSim 2009.

8. M. H. Radfar, R. M. Dansereau, and A. Sayadiyan (2007), "Monaural speech segregation based on fusion of source-driven with model-driven techniques", Speech Communication, 49(6), pp 464-476.

9. P. David (2002), "Experiments with speaker recognition using GMM," in Proc. of Radioelektronika 2002, Bratislava, pp 353–357.

10. Tzu-Chuen Lu, Ching-Yun Chang (2010), "A survey of VQ codebook generation", Journal of Information Hiding and Multimedia Signal Processing, 1(3), pp 190-203.

11. Lloyd, S (1957), least squares quantization in pcm. Bell Telephone Laboratories Paper, Marray Hill.

12. Wallace, R (1989), finding natural clusters through entropy minimization. Ph.D Thesis. Carnegie-Mellon Uiversity, CS Dept.

13. Chris Ding and Xiaofeng He (2004), "K-Means Clustering via Principal Component Analysis", In proceedings of the 21st International Conference on Machine Learning, Banff, Canada.

14. D. Napoleon, S. Pavalakodi (2011), "A New Method for Dimensionality Reduction using K-Means Clustering Algorithm for High Dimensional Data Set" International Journal of Computer Applications (0975 – 8887) 13(7).

15. M. H. Radfar, R. M. Dansereau, and A. Sayadiyan (2007), "A maximum likelihood estimation of vocal-tract-related filter characteristics for single channel speech separation," EURASIP Journal on Audio, Speech, and Music Processing, no. 84 186.

16. Piyush Lotia, M. R. Khan (May, 2011), "Multistage VQ based GMM for text independent speaker identification system", IJSCE, 1(2)

17. P. Li, Y. Guan, S. Wang, B. Xu, and W. Liu (2010), "Monaural speech sep- aration based on MAXVQ and CASA for robust speech recognition," Computer Speech and Language, 24(1), pp 30–44.

18. P. Prabhu, N. Anbazhagan (2011), "Improving the performance of k-means clustering for high dimensional data set", International Journal on Computer Science and Engineering (IJCSE), 3(6).

19. ITU-T Rec P.862 (2001), "Perceptual Evaluation of Speech Quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs", international telecommunication union, Geneva, Switzerland.

20. Milos Cernak, Milan Rusko (2005), "An evaluation of synthetic speech using the PESQ masure", Forum Acusticum, Budapest.