An Unsupervised Approach to Co-channel Speech Separation

Pilla mohan ganesh [1], **M Jaganmohan Reddy** [2], **Naresh Jeggumanthri** [3]

[1,2,3] *Asst.Prof,Computer Science Department,* Vignan's institute of Engineering for women ,A.P,INDIA

***Abstract*— Speech separation and its recognition is based on two different phenomenons, first is speech separation and second is speech recognition. The speech separation is based on the time-domain which is depends on the full unconstrained decomposition of the speech sample just because in constrained approach it become practically hard to compute and also limits the performance of the system. The decomposition is done by an appropriate independent component analysis (ICA) algorithm giving independent components that are grouped into clusters corresponding to the original sources. Speech recognition (SR) is aimed to recognize the speech in large population. And in large population is very time-consuming and impose a bottleneck. So for fast recognition we use GMM based *k*-mean algorithm for fast recognition of speech. For speeding up the whole process the clustered signals are used. Then during the test stage only a small proportion of speaker models in selected clusters are used in the likelihood computations resulting in a significant speed-up with little to no loss in accuracy.**

***Keywords*- Speech separation, Speech recognition (SR), Clusters, GMM, Independent component analysis (ICA)**

I. INTRODUCTION

Blind Signal Separation is the general problem of determining original sources when only their mixtures are available for observation. Over the past 5 years, research on this topic has exploded due to the emergence of relatively successful separation algorithms, as well as the growing sentiment that the technique constitutes a universal panacea capable of everything from de-noising speech to uncovering the laws of the stock market.

The process is often termed "blind", with the understanding that both source signals and mixing procedure are unknown [1]. Such a statement is of course blatant exaggeration –indeed the assumption of *some* specific mixing model is the paramount piece of prior information required, and in many scenarios even knowledge of certain source statistics is necessary.
We thus begin with the channel model:

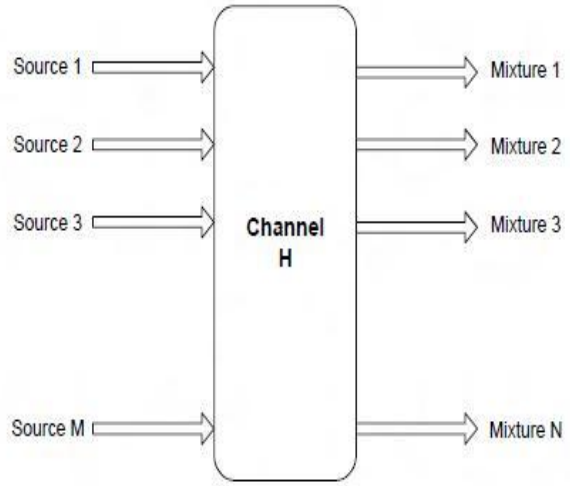


Figure 1: Block Diagram of the Mixing Model

The sources may be sounds, images, biomedical or financial data. Our primary interest will be in audio source signals, with microphones to collect the output mixed signals.

Under this setting, the channel **H** may generally be construed as a linear time-invariant (LTI) system, though there is some activity occurring with nonlinear mixing models.

Three levels of complexity are discerned:

- **H** is a matrix. We call this the *instantaneous mixing model*, since only the relative attenuations of sound due to the microphone-source distances are accommodated.
- $H_{ij} = a_{ij}z^{D}_{ij}$. This is the *delayed mixing model*, incorporating not only the attenuation $\boldsymbol{a_{ij}}$ between the $\boldsymbol{i^{th}}$ microphone and $\boldsymbol{j^{th}}$ source, but the travel time $\boldsymbol{d_{ij}}$ as well. A matrix of FIR filters $\boldsymbol{H_{ij}} = \sum$ **.** This is the *convolutive mixing model,* where room reverberation is accounted for.

Further generalization admits a non-dimension preserving **H:** $N \neq M$. Another attempt at realism introduces a dynamic environment equipped with moving speakers: $\mathbf{H} = \mathbf{H}(t)$. Finally, we may include microphone (sensor) noise $\mathbf{n}(t)$ with the model, though it is possible to consider noise as an additional source. For the latter reason we do not deal with noise in this report; however, methods are available for estimating and eliminating such deteriorating effects without analyzing them as sources.

Though many papers purport to introduce new methods of solution, the existing framework (and solutions) for blind signal separation are often the same. Sources are modeled as random processes despite their essential deterministic mode of production1, and the statistical independence of these random sources is then exploited2. Specifically, the determining criterion for separation is a measure of independence, typically represented by some cost function *J*. The extremum of *J*, with respect to the parameters of some inverse mixing process, then corresponds with more or less independent outputs.

Algorithms which rely on this concept, the separation-independence equivalence, may be classed as those performing *Independent Component Analysis*. The problem of blind signal separation is then reduced to a mathematical *optimization* problem, upon which a multitude of tested techniques may be brought to bear.

The whole paper is divided into four main parts in first part the introduction of the topic is given in second part it is the literature survey and in final the conclusion of the paper is given.

II. Literature Survey

The novel time-domain algorithm has been proposed for blind separation of audio sources in [1] that is based on the complete unconstrained ICA decomposition of the observation space. The algorithm, named T-ABCD, is suitable for situations where only short data records are available. In this respect, it outperforms other known time-domain BSS algorithms. T-ABCD consists of five steps, each one providing a room for other variants and improvements. In particular, the selection of eigenmodes may lead to a more effective definition of the observation space. The comparison with the oracle algorithm showed that the measure of the similarity of components, their clustering, and weighting might be still significantly improved.

In [3] a new single-channel speech-separation approach based on modulation-frequency detection and cross-channel correlation of instantaneous frequency is presented and evaluated.

Using two different databases, we demonstrated that separation using instantaneous modulation frequency provided better recognition accuracy than separation based on fundamental frequency alone and baseline processing using MFCC features. We expect to observe further improvements by incorporating more accurate speaker identification and methods for dealing with unvoiced segments.

In [4] compare speech separation performance of a number of algorithmic and ground truth masks and examined a number of metrics for evaluating the performance of such masks. Automatic speech recognition results should be the best predictor of intelligibility, but we did not find that ASR results for any of the masks under evaluation predicted important aspects of human intelligibility particularly well. Specifically, while the numerical ASR results for ground truth separations followed human intelligibility in anechoic conditions, they were much better than human performance in reverberation. Algorithmic separations, on the other hand, were also similar to human performance in anechoic conditions, but much worse than human performance in reverberation. These results suggest that the separation algorithms investigated in this paper cannot reject reverberant energy as well as humans can.

In [5] it presents comparative assessment of Blind Source Separation methods for instantaneous mixtures. The study highlights the underlying principles of Principal Component Analysis (PCA) and Independent Component Analysis (ICA) in this context. These methods have been tested on instantaneous mixtures of synthetic periodic signals, monotonous noise from electromechanical systems and speech signals. In particular, methods based on Nonlinear PCA, Maximum Entropy, Mutual Information Minimization and Fast ICA, have been compared for their separation ability, processing time and accuracy. The quality of the output, the complexity of the algorithms and the simplicity(implementation) of the methods are some of the performance measures.

In the complete ICA network, it may be advisable to use neural learning only for the vital separation process. For whitening and estimation of basis vectors standard numerical methods will suffice. The poor performance in terms of computation time and separation results may be attributed to the presence of noise and statistical dependency amongst the original source signals. The separation results could probably be improved by adding to the basic network structure, another separating stage where a different nonlinearity may be employed to introduce higher order statistics.

A new technique for the blind separation of convolutive mixtures is proposed in[6]. Start from the application of Kullback–Leibler divergence in frequency domain, and then we integrate Kullback–Leibler divergence over the whole frequency range of interest to yield a new objective function which turns out to be time-domain variable dependent. In other words, the objective function is derived in frequency domain which can be optimized with respect to time domain variables. The proposed technique has the advantages of frequency domain approaches and is suitable for very long mixing channels, but does not suffer from the local permutation problem as the separation is achieved in time-domain.

A polar-coordinate activation function was exploited for complex-valued signals. The objective function was minimized with respect to the channel parameters of the separation system, and the corresponding algorithm was developed. The local frequency-domain permutation problem was avoided through the frequency-domain integration and time-domain optimization[6].

In [7] the performance of several source separation systems on a speech separation task for which human intelligibility has previously been measured. For anechoic mixtures, automatic speech recognition (ASR) performance on the separated signals is quite similar to human performance. In reverberation, however, while signal separation has some benefit for ASR, the results are still far below those of human listeners facing the same task. Performing this same experiment with a number of oracle masks created with *a priori* knowledge of the separated sources motivates a new objective measure of separation performance, the Direct-path, Early echo, and Reverberation, of the Target and Masker (DERTM), which is closely related to the ASR results. This measure indicates that while the non-oracle algorithms successfully reject the direct-path signal from the masking source, they reject less of its reverberation, explaining the disappointing ASR performance.

In [8] the Co-channel (two-talker) speech separation is predominantly addressed using pertained speaker dependent models. In [8] they proposed an unsupervised approach to separating Co-channel speech. This approach follows the two main stages of computational auditory scene analysis: segmentation and grouping. For voiced speech segregation, the proposed system utilizes a tandem algorithm for simultaneous grouping and then unsupervised clustering for sequential grouping. The clustering is performed by a search to maximize the ratio of between- and within-group speaker distances while penalizing within-group concurrent pitches.

To segregate unvoiced speech, we first produce unvoiced speech segments based on onset/offset analysis.

The segments are grouped using the complementary binary masks of segregated voiced speech. Despite its simplicity, our approach produces significant SNR improvements across a range of input SNR. The proposed system yields competitive performance in comparison to other speaker-independent and model-based methods.

The performances of different techniques are as shown by the diagrams for different signal separation and recognition.

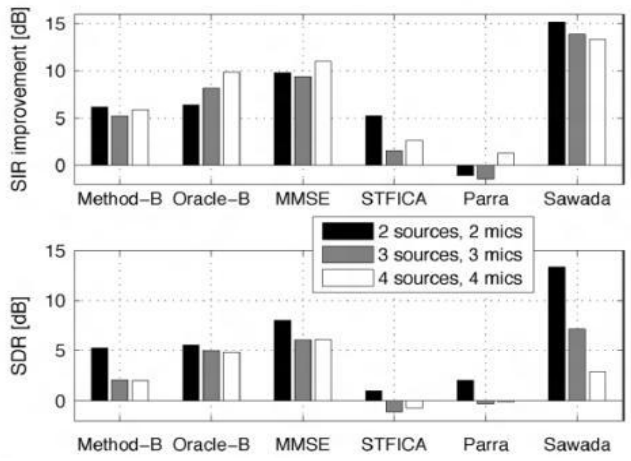
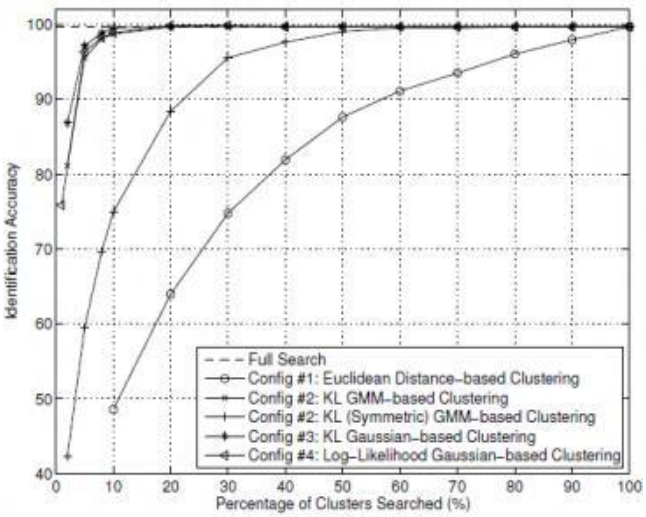
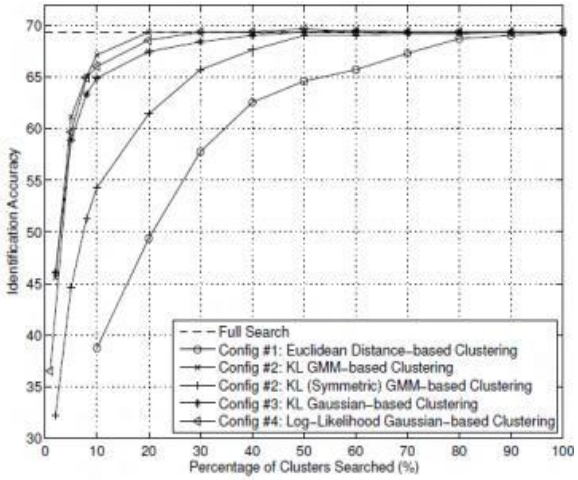


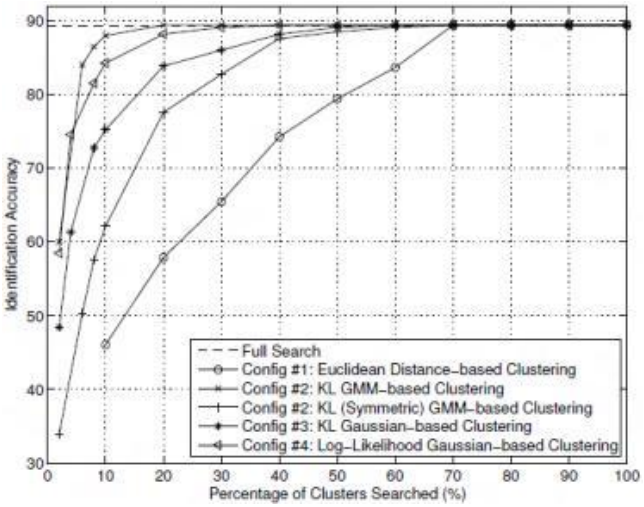
Fig.1. Performance of separation methods applied to Sawada's data with 2–4 sources and microphones.



(a) TIMIT



(b) NTIMIT



(c) NIST 2002

Fig. 2. Speaker identification accuracy versus percentage of clusters searched for (a) TIMIT, (b) NTIMIT, and (c) NIST 2002.

III. CONCLUSION

There are different approaches are proposed for the blind source separation and the speaker identification. The Time-Domain separation method achieved good performance for short data record by applying T-ABCD algorithm. And also in field of speaker recognition a lot of methods were employed but using GMM-based cluster model the complexity of recognition in crowed has been reduced.

During the test stage only a small proportion of speaker models in selected clusters are used in the likelihood computations resulting in a significant speed-up with little to no loss in accuracy. By the application of the TIMIT, NTIMIT, and NIST 2002 corpora search as few as 10% of the speaker model space and realize an actual speed-up of 8:7× with only a small loss in accuracy.

REFERENCES

[1] "Time-Domain Blind Separation of Audio Sources on the Basis of a Complete ICA Decomposition of an Observation Space" Zbyn˘ek Koldovský, Member, IEEE, and Petr Tichavský, Senior Member, IEEE, IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 0, NO. 0, 2010

[2] "Speaker Model Clustering for Efficient Speaker Identification in Large Population Applications", Vijendra Raj Apsingekar and Phillip L. De Leon, Senior Member, IEEE, IEEE TRANS. AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 17, NO. 4, MAY 2009

[3] "SINGLE-CHANNEL SPEECH SEPARATION BASED ON MODULATION FREQUENCY", Lingyun Gu and Richard M. Stern, Language Technologies Institute Department of Electrical and Computer Engineering Carnegie Mellon University,5000 Forbes Ave, Pittsburgh, PA, 15213, U.S.A

[4] "Evaluating Source Separation Algorithms With Reverberant Speech", Michael I. Mandel, Member, IEEE, Scott Bressler, Barbara Shinn-Cunningham, and Daniel P. W. Ellis, Senior Member, IEEE, IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 18, NO. 7, SEPTEMBER 2010

[5] "ICA Methods for Blind Source Separation of Instantaneous Mixtures: A Case Study", Niva Das Department of Electronics and Telecom. Engineering, I.T.E.R. Bhubaneswar India, **Aurobinda Routray** Department of Electrical Engineering, I.I.T. Kharagpur India, Pradipta Kishore Dash Director C.O.E. Bhubaneswar India, Neural Information Processing – Letters and Reviews Vol. 11, No. 11, November 2007

[6] "Blind Source Separation Based on Time-Domain Optimization of a Frequency-Domain Independence Criterion", Tiemin Mei, Jiangtao Xi, Member, IEEE, Fuliang Yin, Alfred Mertins, Senior Member, IEEE, and Joe F. Chicharo, Senior Member, IEEE, IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 14, NO. 6, NOVEMBER 2006

[7] "Evaluating Source Separation Algorithms With Reverberant Speech", Michael I. Mandel, Member, IEEE, Scott Bressler, Barbara Shinn-Cunningham, and Daniel P. W. Ellis, Senior Member, IEEE, IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 18, NO. 7, SEPTEMBER 2010

[8] "An Unsupervised Approach to Co-channel Speech Separation", Ke Hu, Student Member, IEEE, and De Liang Wang, Fellow, IEEE, IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 21, NO. 1, JANUARY 2013

Pilla mohan ganesh working as a Asst professor in Vignan's institue of Engineering for womens in Dept of Information Technology. I obtained **Mtech** Degree from JNTUK .and I obtained **MCA** degree from **Andhra university Campus**, Vizainagarm. I have 3 years of teaching experience in this Field.

M Jaganmohan Reddy working as Asst professor in vignan's institue of engineering for womens in Dept of Computer science. I got **Mtech** degree from JNTUK i have 5 year of teaching experience.

Naresh Jeggumanthri working as a Asst professor in Vignan's institute of Engineering for women in Dept of Information Technology. I obtained **Mtech** Degree from JNTUK .and I obtained M.sc degree from **GITAM** I have 2 years of teaching experience in this Field.