IJIREEICE

ISSN (Online) 2321 – 2004
ISSN (Print) 2321 – 5526

International Journal of Innovative Research in
Electrical, Electronics, Instrumentation and Control Engineering

ISO 3297:2007 Certified

Vol. 4, Issue 9, September 2016

Survey of Methods and challenges in Computational Auditory sense analysis

V.A. Mane[1], Prof. Dr. S. B. Patil[2]

Department of Electronics & Telecommunication, A.D. College of Engineering Ashta[1]

Department of Electronics & Telecommunication, J.J. Magdum College of Engineering, Jaysigpur[2]

Abstract: Sound is a primary form of communication between human beings. Sound coming to our ears is a mixer of sounds from various sources. It is ability of human hearing capability to extract a sound of particular source while rejecting the sounds from other sources. Human ear can do this effortlessly but how machine can do this?. This problem is noted as cocktail party problem. In this paper it is study of various literature and methods used to solve cocktail party problem.

Keyword: CASA (Computational Auditory Sense Analysis), MVDR (Minimum Variance Distortion less Response), Ideal Binary Mask, cocktail party.

I. INTRODUCTION

Every day we listen sound from various sources like speeches from crowd, various vehicle noise, wind noise, voice of friend sitting on bike, crowd and many more. This mixer of sound reaching to our ears from various sources. Now it is ability of human and non human leaving animals to extract a sound of a particular speaker from mixer of various sounds. Researchers are trying to define a process of this human auditory system since several decades and trying to apply it in machine learning. Although humans, and nonhuman animals, perform sense analysis with apparent ease, how machines can extract a sound of a interest while rejecting sound from other sources effortlessly.

In 1953 E.C.Cherry[9] noted this problem as cocktail party problem. In his published paper on "Some experiments on the recognition of speech, with one and with two ears" he presented various experiments carried out, based on listening by on ear and by two ears. The attempt maid by paper is to understand the process of human auditory system of extracting sound of interest while rejecting the other sounds. The first set of experiment carried out, relates to this general problem of speech recognition. How do we recover what one person is saying when others are speaking at the same time. And if it is required to incorporate such a system in machine on what logical basis one can design a machine? One of the logic may be a) Voices from different directions b) lip reading gestures c) different speaking voices, mean pitches, mean speeds, male and female etc. d) Accent differing e) Transition probability. Among all five, last logic of transition probability cannot be excluded. Because human brain may have large set of transition probabilities on which it may enables to predict a particular sound and source of sound with maximum like hood estimation. Some people object on storage of probabilities in brain. Then the question remains the same on what logic we can use to design a machine which will analogous to human being. The test carried out by E.C.Cherry purport to show that human is having such power based on the probabilities ranking of words, phonemic, syntactical ending and other factors of speech and sound. To find out mechanism of human auditory system the experiment is presented with two mixed speeches recorded on a tape, and is asked to repeat one of the speaker voice word by word and phrase by phrase. One can play a tape as many times as he wish without writing it down. An Experimental result showed that less errors occurs in repetition of same tape and hearing the same sound number of times. Improvement in playing words and phrases seen after repeating the tape for more number of time. Now the same experiment is carried out with writing it down and now errors are seen minimum. Also except some grammatic mistakes the long phrase have identified correctly. Another set of experiment carried out is related with unmixed speeches. At this time two different messages were recorded by same speaker. Now this is played one in the left ear and other in the right ear for observation purpose how human auditory system interact with this. It has been obseved that Speaker have not found any difficulty in listening and understanding any one of the message from any one of the ear as it is a natural behaviour of human to reject unwanted speech similar to, if any one tries to listen conversation of speakers in crowd, sudden action takes place to turn on one ear towards conversation. And among the conversation also on the interested person or conversation. Now for speaker if it is asked about what he listened other than conversation then most of time reply is crowd noise. Human auditory system listens everything but extract and concentrate on sound of interest [9].

In another experiment two different messages were started in both ears in English spoken by one speaker. When listener was concentrating on right ear, suddenly if the

IJIREEICE

ISSN (Online) 2321 – 2004
ISSN (Print) 2321 – 5526

International Journal of Innovative Research in
Electrical, Electronics, Instrumentation and Control Engineering

ISO 3297:2007 Certified

Vol. 4, Issue 9, September 2016

language in left ear is changed to German for some time span but spoken by the same speaker then it is found that listener will reject that he had listened German voice also. Because he did not know rejected message.

Shannon has already reported that prediction is readily possible in case of printed language [9]. It is possible to decode a written message of a particular person from mixer of written message by observing combined message. It is possible to decode the message based on successive identification words, writing style of letters etc., and then grouping the words to form a whole sentence. But it is quite difficult to segregate speech or sound of particular source from a mixer by machine, even though our ears can do this effortlessly. Sins then it become a matter of interest to so many researchers to define an underlying process of human listening capability of sound separation and segregation from same source of sound. Following are some methods and and algorithms made in the field of Computational Auditory Sense Analysis (CASA) in various applications.

II. ASA (AUDITORY SENSE ANALYSIS)

Bregman [15][14] was the first to present a logical answer to the of cocktail party problem. He contends that listeners perform an auditory scene analysis (ASA), which can be conceptualized as a two-stage process. In the first stage, the acoustic mixture is decomposed into elements. An element may be regarded as an atomic part of the auditory scene, which Computational Auditory Scene Analysis describes a significant acoustic event. Subsequently, a grouping process combines elements that are likely to have arisen from the same acoustic source, forming a perceptual structure called a stream.

In the article written by Challenges for Computational Intelligence, W. Duch and J. Mandziuk (Eds.), Springer publication [14], it is explained about human auditory system. Human auditory system is based on perception, action and reasoning shown in figure 1.

Figure 1. Perception, reasoning and action

Perception is just a sense. Senses can be a auditory, visual, touch, smell etc. For ASA (Auditory Sense analysis), sense is auditory. Perception is an input to reasoning and Action takes place on perception. Reasoning plays an important role of doing action based on perception. Reasoning is the process what human brain can do. In other words, perception and action are about input and output, from the viewpoint of the intelligent agent (i.e. a human being). Reasoning involves higher cognitive functions such as memory, planning, language understanding, and decision making, and is at the core of traditional artificial intelligence. Reasoning also serves to connect perception and action, and these three aspects interact with one another to form the whole of intelligence. Although humans, and nonhuman animals, perform scene analysis with apparent ease, computational scene analysis remains an extremely challenging problem despite decades of research in fields such as computer vision and speech processing. To understand perceptual information processing by machine, it requires three different levels of descriptions. The first level of description, called computational theory, which is mainly concerned with the goal of computation. The second level, called representation and algorithm,(representation concerned with representation of the input and the output. and the algorithm that transforms from the input representation to the output representation) The third level, called hardware implementation(concerned with how to physically realize the representation and the algorithm). The goal of computational scene analysis is to produce a computational description of the objects and their spatial locations in a physical scene from sensory input. The above goal of computational scene analysis is strongly related to the goal of human scene analysis. Progress in computational scene analysis may shed light on perceptual and neural mechanisms[14]. Various researchers have developed the methods and applications for using Computational Auditory Sense Analysis (CASA) frame work as a front end for many applications. Some of the applications and methods presented here from various authors are found as below.

III. MONAURAL SYSTEM

Monaural speech separation [8] is a challenging problem in speech and signal processing. Monaural speech separation system works as shown in figure 2. The sound from source A and B are recorded using single microphone. Sound or speech separation system can able to separate the sound source A and B.

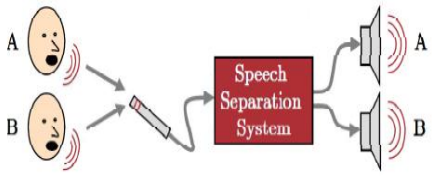

Figure 2: Speech separation system

This Speech segregation, also known as the cocktail party problem, refers to the problem of segregating target speech from its background interference [7][2]. Monaural speech segregation, attempt speech/sound segregation by monaural recordings with one microphone. Recoding by one microphone may be sufficient to design mechanism like human auditory system. This is important for many real-world applications including robust speech and speaker

IJIREEICE

ISSN (Online) 2321 – 2004
ISSN (Print) 2321 – 5526

International Journal of Innovative Research in
Electrical, Electronics, Instrumentation and Control Engineering

ISO 3297:2007 Certified

Vol. 4, Issue 9, September 2016

recognition, audio information retrieval and hearing aids design and in artificial intelligence [8]. However, despite decades of effort, monaural speech segregation still remains one of the challenging problems in signal processing. Numerous algorithms have been developed to solve the monaural speech segregation problem. For example, spectral subtraction [13] and Weiner filtering are two representative techniques [4]. The auther Peng Li, Yong Guan, Bo Xu, and Wenju Liu "Monaural Speech Separation Based on Computational Auditory Scene Analysis and Objective Quality Assessment of Speech" in his paper, proposed method mainly focused on performance evaluation of CASA signal-to-noise ratio (SNR) improvement. They proposed a new method which combines CASA with objective quality assessment of speech (OQAS). With this combination, the performance of the speech separation can be improved not only in SNR, but also in mean opinion score (MOS). The CASA system employed in the proposed model is based on the Hu and Wang's model [4]. This model is a typical primitive CASA model used to segregates resolved and unresolved harmonic. Segregation of resolved harmonics is based on temporal continuity and cross-channel correlation. The system generates segments and groups them according to their periodicities. And unresolved harmonics segregation, generates segments based on common amplitude modulation (AM) in addition to temporal continuity and groups them according to the AM rates [4]. To get an accurate pitch contour, a coarse one is first estimated with speech segregated according to the dominant pitch. Then it is revised according to psychoacoustic constraints. This makes the separation performance of the Hu and Wang's model almost best in the primitive CASA systems and even better than many other knowledge-based CASA systems in dealing with voiced speeches [4].

In Hu and Wang's model selected by author employs the notion of time-frequency mask or Ideal Binary Mask (IBM). The idea of binary mask is supported by the auditory masking phenomenon within a critical band. A weaker signal is masked by a stronger one in case certain band of frequency. The ideal binary mask is very effective for human speech intelligibility. It provides an excellent front end for automatic speech recognition (ASR). It also provides a convenient way to combine the CASA system with the OQAS algorithm. The author has proved that the proposed model can effectively improve the SNRs and most perceptual qualities of the separated speeches. In comparison of the proposed algorithm with other separation or enhancement systems, it had drawn a conclusion that the proposed method is effective in processing the monaural speech separation problem.

Voice activity detection (VAD) is another application of CASA, widely used in the automatic speech recognition (ASR) and also in mobile communications for the control of discontinuous transmission schemes and in many noise tracking algorithms for speech enhancement [2]. The accuracy of VAD system is becoming more demanding with the development of speech applications in real environment. However, VAD is still facing a problem with the presence of background noise, especially in non stationary noise. A typical VAD system proposed by author [2] contains mainly two parts: feature extraction and decision making. For feature perspective, VAD systems often employ time-domain features such as energy and zero crossing rate, spectral-domain features such as spectral difference, Discrete Fourier Transform (DFT) coefficients, cepstral-domain features such as Mel-frequency cepstral coefficients (MFCC), harmonicity-basd features such as the harmonic structure-based VAD features, the harmonicity of DFT etc. From the decision making perspective, algorithms include thresholding method such as statistical model based methods and machine learning based. The paper focuses on two novel features for VAD based on computational auditory scene analysis (CASA). The first method is based on statistical model based VAD and the second is a supervised method based on Gaussian Mixture Model. In the proposed method gammatone frequency cepstral coefficients (GFCC) are extracted from cochleagram instead of DFT coefficients and used these feature to discriminate speech and noise in noisy signal. Gaussian mixture model is also used to model GFCC of speech and noise. The performances of the proposed methods are compared with several existing algorithms. The results demonstrated that CASA based features outperform several traditional features in the task of VAD. Result have been evaluated based on TIMIT database and proved that GFCC extraction is better than MFCC.

Above literature focus on segrigation of voiced speeches. Lot of effort has been made in computational auditory scene analysis to segregate voiced speech from monaural mixtures but much attention is not received for unvoiced speech segregation. Unvoiced speech is highly susceptible to interference due to its relatively weak energy and lack of harmonic structure, and hence makes its segregation extremely difficult[13]. This paper[13] proposes a new approach to segregation of unvoiced speech from non speech interference. The proposed system first removes estimated voiced speech, and the periodic part of interference based on cross-channel correlation. And estimate the noise energy in unvoiced intervals using segregated speech in neighboring voiced intervals. Unvoiced speech segregation occurs in two stages: segmentation and grouping. In segmentation, author applied spectral subtraction to generate time–frequency segments in unvoiced intervals. Unvoiced speech segments are subsequently grouped based on frequency characteristics of unvoiced speech using simple thresholding as well as Bayesian classification. The proposed algorithm is computationally efficient, and systematic evaluation and comparison showed that this approach considerably improves the performance of unvoiced speech segregation [13]. The basic idea of unvoiced speech segregation method proposed is to capitalize on the segregated voiced speech to estimate interference energy. Estimated voiced binary mask

IJIREEICE

ISSN (Online) 2321 – 2004
ISSN (Print) 2321 – 5526

International Journal of Innovative Research in
Electrical, Electronics, Instrumentation and Control Engineering

ISO 3297:2007 Certified

Vol. 4, Issue 9, September 2016

contains inactive T-F units during voiced intervals, this has been used to estimate noise energy and subtract it from the mixture during unvoiced intervals in order to form unvoiced segments. And then removing periodic signals. They estimated the background noise and then removed it during unvoiced intervals. And then estimate the interference energy in an unvoiced interval by averaging the mixture energy within inactive T-F units in the two neighboring voiced intervals. The proposed method uses 64-channel gammatone filterbank in T-F analysis Compared with systems employing 128-channel filterbanks the use of a 64-channel filterbank halves the computing time also.

IV. CHALLENGES IN CASA BASED SYSTEMS

It has been studied extensively, and many separation systems based on computational auditory scene analysis (CASA) have been proposed in the last several decades. The speech extraction systems based on Monaural Speech Separation based on Computational Auditory Scene, thinks over one sensor is sufficient to understand auditory sense in the application where location of sound is not so important. Even use of only CASA is not sufficient. The segregated signal may get captured by various noises. There is need to improve the signal to noise ration also. Combining Object quality algorithm with CASA improves signal to noise ratio for CASA based applications. Although the research on CASA has tended to introduce high-level knowledge into separation processes using primitive data-driven methods. The knowledge on speech quality still has not been combined with it. By combining Object quality algorithm and CASA, improves the signal-to-noise ratio (SNR). Even tough, the quality of the separated speech is not directly related to its SNR.

The main challenge in CASA based system is a method used to separate the sound sources and segregation of sound from same source. Classification of sound based on speech segregation Monaural speech segregation has been a problem for several decades [4][13]. By casting speech segregation as a binary classification problem, advancement has been made in computational auditory scene analysis on segregation of both voiced and unvoiced speech. So far, pitch and amplitude modulation spectrogram have been used as two main kinds of time-frequency (T-F) unit level features in classification. Better advantage can get by expanding T-F unit features to include gamma tone frequency cepstral coefficients (GFCC), mel-frequency cepstral coefficients, relative spectral transform (RASTA) and perceptual linear prediction (PLP).

Another method used for segregation of speech is Ideal binary mask [1]. This method is used occasionally to separate speech of interest from various sound sources. Similar to human auditory scene analysis, computational auditory scene analysis (CASA) approaches the segregation problem on the basis of perceptual principles. A commonly used computational goal in CASA is the ideal binary mask (IBM), which is a two-dimensional matrix of binary labels 1 and 0. Where 1 indicates that the target signal dominates the corresponding time-frequency (T-F) unit and 0 otherwise. By masking target speech from other sound it is possible to separate and segregate target speech. Recent speech perception research shows that IBM segregation produces large improvements of speech intelligibility in noise for normal-hearing listeners and hearing-impaired listeners. Such improvements persist when room reverberation is present. Even though human listening system can separate to recognize and segregate target sounds. Conventional automatic speech recognizer does not perform well in the presence of multiple sound sources. Computational auditory scene analysis system for separating and recognizing target speech in the presence of competing speech or noise can be effective with use of Ideal Binary Mask. The ideal binary time–frequency (T–F) mask which retains the mixture in a local T–F unit if and only if the target is stronger than the interference within the unit.

Recognizing a speech in robust environment faces main challenges. Hoang Do, and Harvey F. Silverman [6] proposed concept of sound source separation Algorithm for an Adverse Environment That Combines MVDR-PHAT with the Casa Framework. Extracting a high-quality speech signal of a single source from a multiple-source input in an adverse environment has always been a challenge for microphone-array processing. Three major approaches have been proposed to tackle this problem. Use of blind-source separation (BSS)[11] for sound source separation, beam forming (BF) for segregation of sound, and computational auditory scene analysis (CASA) can improve the quality of application. Combinations of the CASA, BSS and BF also have been introduced. They proposed a new algorithm which utilizes the null-steering beam former minimum variance distortion less response (MVDR) using the proven-robust phase transform (MVDR-PHAT) and the CASA framework that closely mimics human hearing perception. Experimental results using real data recorded in a room with high background and reverberation noise indicated the improved performance of the proposed algorithm compared to those of traditional beam forming algorithms and an SRP-PHAT-based source-separation algorithm.

The basic idea of the algorithm is:

i. A time-frequency masking of a delay-sum beam former (DSBF) steered to the hypothesized point-source location of the desired source is derived.

ii. An MVDR-PHAT power spectrum is used as the functional to compute a value for each (t, f).

iii. (t, f) points with appropriately high functional values compared to that of background noise level are maintained, whereas low functional valued points will be suppressed.

iv. An inverse short-time Fourier transforms (ISTFT)

Speech data from two fixed sources (loudspeakers) recorded using the Huge Microphone Array (HMA) [6] in

IJIREEICE

ISSN (Online) 2321 – 2004
ISSN (Print) 2321 – 5526

International Journal of Innovative Research in
Electrical, Electronics, Instrumentation and Control Engineering

ISO 3297:2007 Certified

Vol. 4, Issue 9, September 2016

a real room with T60 = 450 ms was used to compare the performance of the proposed algorithm to those of a delay-sum beam former (DSBF), a minimum variance distortion less response beam former (MVDR), and the recently proposed CASA using SRP-PHAT. Experimental results showed the proposed algorithm performed the best of the three algorithms. The result obtained by the author is presented in table 1 below [6]. Shows the segSNR and PESQ measures for source1.

TABLE1: segSNR and PESQ SCORE FOR SOURCE1

Algorithm	segSNR	PESQ
DSBF	-6.0403	2.012
MVDR	-6.1517	1.0796
CASA+SRP+PHAT	-4.7636	1.9052
CASA+MVDR+PHAT	-4.3379	1.9759

Author shown that using CASA+MVDR+PHAT improve SNR and PESQ score. Another result found with source2 for same parameters are given in table 2 below [6].

TABLE2. segSNR and PESQ SCORE FOR SOURCE2[6]

Algorithm	segSNR	PESQ
DSBF	-6.1267	1.9678
MVDR	-6.2086	1.9490
CASA+SRP+PHAT	-4.4318	2.0871
CASA+MVDR+PHAT	-4.3379	1.9759

Algorithm measuring the segSNR measure for isolating each of the desired sources. The results are somewhat mixed for the PESQ measure, Perhaps the PESQ measure is more meant for quality of coders in a non-interferer environment. The three measures of the most successful baseline (CASA + SRP-PHAT) and the proposed algorithm are presented in table 3 below [6].

TABLE 3 SDR, SIR, and SAR measures of the two PHAT-based separation algorithms for source 1

Algorithm	SDR(DB)	SIR(DB)	SAR(DB)
CASA+SRP-PHAT	-62.1049	-13.3950	-48.5155
CASA+MVDR-PHAT	-41.8893	-4.1650	-40.4802

Above table shows distortion in SDR (Source-to-distortion ration), source-to-interference ratio (SIR), and source-to-artifact ratio (SAR). Above result of proposed algorithm created less distortion and artifacts than the baseline, while suppressing the interferer more effectively [6].

Robust automatic speech recognition (ASR) in noisy and reverberant environments remains a challenging problem since last decade [10], while considering the relative performance of ASR systems and human listeners on the same speech recognition task. Recognizing a word error rates for ASR systems can be greater than those for human listeners. Differences are largest when the speech is contaminated by background noise or room reverberation. So there is need to implement auditory system that give rise to underlying mechanisms be incorporated machine in order to improve their robustness.

CASA-Based Robust Speaker Identification [5][12] based methods used in traditional way in conventional speaker recognition systems perform poorly under noisy conditions. Like human auditory perception, computational auditory scene analysis (CASA) typically segregates speech by producing a binary time–frequency mask. Investigated CASA for robust speaker identification, to deal with noisy speech, it is possible to apply CASA separation and then either reconstruct or marginalize corrupted components indicated by a CASA mask. To deal with noisy speech, author [5] has applied CASA separation and then either reconstruct or marginalize corrupted components indicated by a CASA mask. Further combining the two methods into a single system based on their complementary advantages, the system achieves significant performance improvements over related systems under a wide range of signal-to-noise ratios**.** A proposed robust speaker identification system by using CASA as a front-end to perform speech segregation uses CASA segregation is in the form of a binary time–frequency (T-F) mask that indicates whether a particular T-F unit is dominated by speech or background noise. GF is obtained from a bank of gammatone filters. Then GFCC is derived from GF by a cepstral analysis. Author showed in that GFCC achieves an SID level of performance in noisy environments that is significantly better than MFCC. Following Table 4 shows SID accuracy (%) of the combined system and baselines performance is averaged across different SNR conditions[5]. Where GFCC_22 represent 22 dimensional GFCC and MFCC.

Table 4 AVERAGE SID ACCURACY

method	babble	factory	Ssn	average
Combined system	72.58	71.33	71.18	71.7
gfcc_22	47.64	51.46	49.61	49.57
mfcc_22	39.42	35.95	31.58	35.65
mfcc_12	35.27	29.7	26.55	30.51
Etsi-afe_d	40.55	45.33	43.27	43.05

The combined system's SID results are more than 28 percentage points higher than those of MFCC and ETSI-AFE_D baselines. Under clean conditions, MFCC_22 yields the SID accuracy of 96.67% (94.39% for MFCC_12), whereas the accuracy is 97.12% for GFCC_22. GF as a spectral feature gives the accuracy of 95.76%, which is slightly worse than the 22-dimensional cepstral features [5]. Another problem in CASA based system is sound localization. Localization of sound in human being is possible with two ears. For mimic the two ears of human being the binaural auditory system is needed. Binaural systems are working just like two ear interface to human. For robotics applications for audio

IJIREEICE

ISSN (Online) 2321 – 2004
ISSN (Print) 2321 – 5526

International Journal of Innovative Research in
Electrical, Electronics, Instrumentation and Control Engineering

ISO 3297:2007 Certified

Vol. 4, Issue 9, September 2016

motor maps the relationship between certain audio features and the position of the sound source is necessary [3]. Binaural methods are used mainly mapping of sound source orientation and distance. Experimental result applied for robotics presented in the paper added auditory sense along with visual sense with CASA framework. It can give remarkable advantages. The method is called as computational audiovisual scene analysis (CAVSA) [3]. It can be adapted online in free interaction with a number of a priori unknown speakers. CAVSA is used to enables a robot to understand real time interaction. In the proposed method it is possible to identify the number and position of speakers, as well as who is the current speaker.

Based on the difference in charactoristics in the recorded signal between the left and right ear, the position of the sound source can be identified. The characteristics used are the interaural intensity difference (IID) and the interaural time difference (ITD) and sometimes spectral components also. IID is the difference in the sound amplitude between the left and right ear (microphone). It makes use of the head-shadow effect, that the head absorbs some of the sound energy and attenuates the sound signal arriving at the far ear especially at higher frequencies [3]. The delay in arrival time of a sound signal between two ears (microphones) (ITD) has similar characteristics to IID. ITD calculated from interaural phase difference is ambiguous at high frequencies. Many different source positions would generate the same ITD value. But IID is only significant at high frequencies, since low frequency signals are not attenuated very much by the head. Both IID and ITD cues needed to perform sound localization in the full range of frequencies [3].

In the work carried out by Rujiao Yan, Tobias Rodemann, and Britta Wrede appled computational audiovisual scene analysis (CAVSA) to search for the visual part of the current sound source for online adaptation of audio-motor maps. CAVSA enables a robot to understand its surroundings when interacting with humans. For example, the robot needs to know how many speakers are present, where the speakers are, and who is currently talking to the robot. In this article, CAVSA is used to link an utterance to a face when robot hear an utterance and see many faces. There is always major role of vision feedback to localise sound source and current speaker. The author has shown that the system was able to bootstrap with a randomized audio-motor map in multi person environments. The number of persons (2–4) has little influence on the learning performance. The percentages of update steps where a wrong person is chosen as the current speaker were about 6% in scenarios with 2–4 persons. Moreover, their system was capable of bootstrapping itself in a natural dialog scenario with 4 human speakers [3].

V. CONCLUSION

Computational Auditory sense analysis can be major step towards mimicry of human auditory system. This can put lot of intelligence in audio based machines especially in the field of robotics. Research in CASA will excel the various fields with real life performance. Above literature tells about development of sense analysis and various steps taken to design and define the underlying process which will do human mimicry.

REFERENCES

[1] Yi Jiang, DeLiangWang, RunSheng Liu, and ZhenMing Feng "Binaural Classification for Reverberant Speech Segregation Using Deep Neural Networks" IEEE/ACM transactions on audio, speech, and language processing, vol. 22, no. 12, pp. 2112-2121, December 2014

[2] Ming Tu, Xiang Xie, Xingyu Na School of Information and Electronics Beijing Institute of Technology "Computational Auditory Scene Analysis Based Voice Activity Detection" 22nd International Conference on Pattern Recognition © 2014 IEEE

[3] Rujiao Yan, Tobias Rodemann, and Britta Wrede "Computational Audiovisual Scene Analysis in Online Adaptation of Audio-Motor Maps" IEEE Transactions On Autonomous Mental Development, Vol. 5, No. 4, pp. 237-287 December 2013

[4] Yuxuan Wang, Kun Han, and DeLiang Wang "Exploring Monaural Features for Classification-Based Speech Segregation" IEEE Transactions On Audio, Speech, And Language Processing, vol. 21, no. 2, pp. 270-279 , February 2013

[5] Xiaojia Zhao, Yang Shao, and DeLiang Wang "CASA-Based Robust Speaker Identification" IEEE Transactions On Audio, Speech, And Language Processing, vol. 20, no. 5, pp. 1608-1916, July 2012

[6] Hoang Do, Harvey F. Silverman EMS School of Engineering Box D, Brown University, Providence, RI 02912, USA "A Robust Sound-Source Separation Algorithm For An Adverse Environment That Combines Mvdr-Phat With The Casa Framework" IEEE Workshop on Applications of Signal Processing to Audio and Acoustics October 16-19, 2011

[7] Yang Shao, Soundararajan Srinivasan , Zhaozhang Jin , DeLiang Wang " A Computational Auditory Scene Analysis System For Speech Segregation And Robust Speech Recognition" Available online 28 March 2008 Elsevier AND Science direct.

[8] Peng Li, Yong Guan, Bo Xu, and Wenju Liu "Monaural Speech Separation Based on Computational Auditory Scene Analysis and Objective Quality Assessment of Speech" IEEE Transactions On Audio, Speech, And Language Processing, vol. 14, no. 6, pp. 2014-2023 November 2006

[9] Cherry, E.C. Some experiments on the recognition of speech, with one and with two ears. J. Acoust. Sot. Am., 25, 975-979, 1953.

[10] Sue Harding, Member, IEEE, Jon Barker, and Guy J. Brown "Mask Estimation for Missing Data" IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 14, NO. 1, JANUARY 2006

[11] Blind Source Separation Exploiting Higher-Order Frequency Dependencies Taesu Kim, Student Member, IEEE, Hagai T. Attias, Soo Young Lee, Member, IEEE, and Te-Won Lee, Member, IEEE IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 15, NO. 1, JANUARY 2007

[12] Application of Shape Analysis Techniques for Improved CASA-Based Speech Separation Yun-Kyung Lee and Oh-Wook Kwon IEEE Transactions on Consumer Electronics, Vol. 55, No. 1, FEBRUARY 2009

[13] Unvoiced Speech Segregation From Nonspeech Interference via CASA and Spectral Subtraction Ke Hu, Student Member, IEEE, and DeLiangWang, Fellow,IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 19, NO. 6, AUGUST 2011

[14] Separation of Speech by Computational Auditory Scene Analysis Guy J. Brown1 and DeLiang Wang2 Reprinted from Speech Enhancement, J. Benesty, S. Makino and J. Chen (Eds.), Springer, New York, 2005, pp. 371–402.

[15] Bregman AS (1990) auditory scene analysis. MIT Press, Cambridge MA

[16] Endeavour, New Series, Volume 17, No. 4, 1993. "Computational auditory scene analysis:listening to several things at once" Martin Cooke, Guy J. Brown, Malcolm Crawford and Phil Green