IOSR Journal of Electronics and Communication Engineering (IOSR-JECE)
e-ISSN: 2278-2834,p- ISSN: 2278-8735.Volume 18, Issue 5, Ser. I (Sept. – October. 2023), PP 07-24
www.iosrjournals.org

Survey Of Algorithms Used In Computational Auditory Scene Analysis For Speaker Identification

Uddhav Shivaji Shid[1,a)], Dr. Suresh D.Shirbahdurkar [2,b)]

[1]*Research Scholar, Department of Electronics and Telecommunication Engineering, Zeal College of Engineering and Research, Pune, Savitribai Phule Pune University, Pune, Maharashtra, India*
[2]*Professor,Zeal College of Engineering and Research, Department of Electronics and Telecommunication Engineering, Savitribai Phule Pune University, Pune, Maharashtra India*

Abstract:
The term "auditory scene analysis" (ASA) refers to the act of breaking down a wide range of acoustic information into discrete auditory perceptual items, such as melodies or underlying physical sources. ASA-related perceptual events are now the subject of a slew of new computational models, several of which have just appeared in peer-reviewed journals. As a result of this review, we hope to connect the theoretical principles of these computational models with the core issues of the framework of theoretical analysis that we have developed. Specific questions include how they achieve grouping and separation of sound elements, as well as whether they apply any type of competition amongst alternate interpretations of the sound input. These theories are examined in terms of the extent to which they incorporate prediction processes, as significant current theories argue that perception is essentially prescriptive. There is a lack of a comprehensive understanding of how the complex acoustic signal is interpreted by existing computational models of ASA, which focus on analyzing the utility of individual procedures (or algorithms) for determining the causes. As a result, a more comprehensive explanation of ASA might incorporate the models' complementing elements.

Keywords: *auditory scenes analysis, ideal binary mask, steady-state suppression, binaural processing*

Date of Submission: 24-09-2023 Date of Acceptance: 04-10-2023

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 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 algorithms made in the field of Computational Auditory Sense Analysis (CASA) in various applications.

II. ASA (AUDITORY SCENSE ANALYSIS)

While attempting to portray the working human auditory system in such a kind of ambiance in the 1950s, Colin Cherry created the term "cocktail party dilemma" (E. C. Cherry 1957) [18]. He conducted a series of tests to determine what elements aid humans have in performing this difficult activity (Brungart P. S et al. 2006) [17]. Since then, a variety of explanations have been advanced to give details of the findings of those tests (W. Speith et al. 1954, E. C. Cherry 1957,D. Brungart el al.2006,) [51] [18][17]. Helmholtz, in 1863[32], had thought on the difficulty of this signal using the model of a ballroom setting in the mid-nineteenth century. Our ears can "identify all the individual constituent parts of this confusing totality," he said, even though the signal is "complex beyond conception." our auditory system deals with the cocktail party phenomenon. In his groundbreaking 1990 book Auditory Scene Analysis [24], Bregman attempted to provide a systematic analysis. By establishing comparisons with eyesight, he names the process "scene analysis." For somebody the perception is used to create mental model of environment. By combining the information gleaned from our senses, our brain constructs mental images of what it has observed. Auditory sense creates the mental representation of an acoustic environment by segregating sound components together, like focusing on the target speaker by suppressing the rest of the sound and cocktail party. According to Bregman, the auditory system carries out this job in two stages. To begin, for each source, separate local time-frequency components are separated. The second stage creates the grouping of those separated elements from each source. This stage is also referred to as segmentation because it creates time-frequency zones (segments) that are locally grouped (D. L. Wang et al. 2006) [58]. The segments from the same source are then grouped together in the second stage to generate an aural stream. A single source is represented as a stream

Bregman solved the problem of the cocktail party by suggesting a realistic solution. He claims that a Listener going through an auditory sense process follows two steps. The first one is an acoustic mixer reaching to ears is broken into pieces. In computer auditory sense analysis, this process is defined as a major and unavoidable part. After this stage, grouping from the same source of sound takes place, which, in other words, forms a stream of signals from each source. In a real sense, the human auditory system constitutes three-part. First, one is perception, second is reasoning, and last is action.

The author Springer edition (Guy J. Brown et al. 2005)[14] has discussed about human auditory system in their essay “Challenges for Computational Intelligence” It has been depicted as in following fig

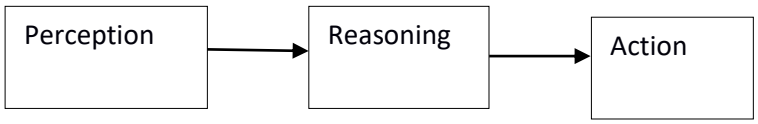

Figure 1.1: Process in Human Auditory system

The senses are the only things that can help us make sense of the world around us. This includes the sense of taste and smell. The senses also include things like hearing and seeing. ASA uses the sense of hearing (Auditory Sense Analysis). It's important to be able to make rational decisions by having a clear picture of what

you're seeing. When it comes to taking action based on their opinion, a person's ability to reason is critical. Reasoning is a process that the human brain is capable of successfully accomplishing. Thus for human beings, perception is about what is seen or heat. These are the building blocks of "classical AI." They include things like memory, planning, and understanding language. Reasoning also connects perception and action, and these three parts of intelligence combine to generate intelligence as a whole. Regardless of decade years of development in domains like computer vision and audio signal processing, computational scene analysis remains a difficult subject despite the seeming simplicity with which humans and nonhuman animals perform it. Three levels of descriptions are required to comprehend perceptual information processing by machines. The computational theory is the initial level of description, and it is primarily concerned with the objective of computing. The hardware development process will come at last to see the physical realization of the developed process.

The goal of computer scene analysis is to use the user's senses to provide a computerized description of the elements and their locations in a real-world scenario. In many ways, the goal of computer scene analysis and human scenario analysis is the same, but they are not the same thing. Computer scene analysis could help with things like perception and neurobiology [14]. This framework, Computational Auditory Sense Analysis (CASA), has been used by many researchers as a front-end for a number of different applications. The following literature is about what various authors talk about this process of auditory sense analysis.

Computational Auditory Scene Analysis

The research of Wang and Brown (2006) [58] carried out experimentation by taking one or two recordings of the acoustic world to get a computer to do what a person would do in ASA. There are two microphones in this area because it is important to biology and to CASA, which is why this specification says that there should be no more than two (as in humans). CASA systems employ perceptually motivated methods. Harmony, for example, is used as a grouping cue in most systems [58]. However, this never implies that the obtained systems be entirely reliant on (Auditory sense analysis) ASA to attain respective objectives. As it has been seen, current systems combine perceptual cues with procedures that aren't always motivated by biological considerations. The goal of ASA is that sound sources be linked to perceptual streams in the auditory information that reaches our ears.

The authors G. Hu et al.2004, G. Hu et al.2001, D. L. Wang et al. 2008 [33][34][57] explained the ultimate goal of CASA. The development of the Ideal Binary Mask for this type of data, according to Wang and others, is one of the primary goals of CASA. The masking phenomena within auditory awareness, stronger (dominant) sound always covers a probably weaker sound and renders an impossible to hear within a critical band (B. C. J. Moore et al.2003)[41]. This became a major inspiration for the concept. In a similar vein, the IBM determines which portions of a mixture's time-frequency representation are target dominant and which are not. The IBM is a binary matrix mask target dominant time-frequency unit is indicated by 1 and interference appearing or dominant time-frequency unit represented as 0. By considering a spectrogram-like representation of an acoustic input reproduced from Wang et al.

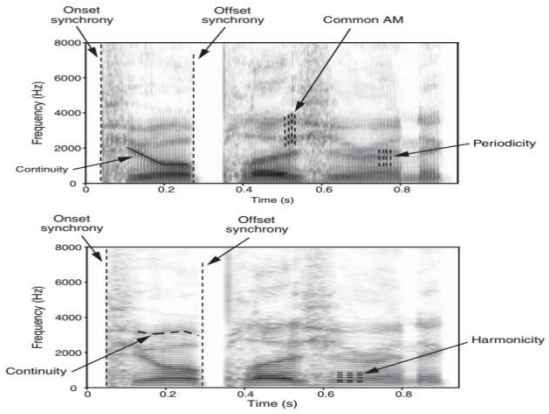

Figure 2.1: Group cues for speech organization

A broadband spectrogram of the phrase "absolute delight" is shown in the upper part of the picture. At the beginning and end, it shows that there is a sense of time and synchronization with amplitude modulation and harmonicas, as well. This is the narrow-band spectrogram of the same phrase. It has been shown in the bottom right corner. In a binary matrix, target dominant T-F units have a single entry, and interference dominant T-F units have no entries. IBM receives this matrix and processes it.

The IBM is mathematically defined as:

$$IBM(t,f) = \begin{cases} 1 & if SNR(t,f) \geq LC \\ 0 & otherwise \end{cases} \quad \text{---------- (2.1)}$$

SNR is the signal-to-noise ratio utilizing the time(t) and frequency(f) indexes(t,f). A unit's signal-to-noise ratio (SNR) must be greater than or equal to the target's. For CASA based model and for ASR required threshold of energy is typically set at 0dB. Pre-mixed target and interference signals are required for the IBM (thus the word "ideal"). These people believe a CASA system can estimate IBM from a mixed-signal. Oracle, a binary mask, is a good metaphor for IBM. To depict the loud speech ceiling recognition performance in missing data ASR experiments, Oracle masks are typically utilized.

Researchers (Y. Li and colleagues, 2009) [39] in his perspective, the following are some of the reasons to explain that IBM is a suitable CASA goal: (i) Li and Wang stated that efficiency of Ideal Binary mask (IBM) can be determined by estimating Signal-to-Noise ration of noisy sample of speech after processing through a processed binary mask. According to the researchers, the IBM with an LC of 0 dB is occasionally the optimal binary mask in certain situations. The ideal ratio (soft) mask and the ideal ratio (hard) mask are two more T-F masks that can be compared to the IBM in terms of performance. According to the measurements, IBM and optimum ratio mask SNR increases are comparable in the majority of interesting mixes. (ii) Both normal and deaf and hard-of-hearing listeners benefit from IBM-segregated noisy speech [20, 17, 38, and 59]. It is possible to increase the intelligibility of noisy speech even if the IBM has tampered with it (N. Li et al.2008) [38]. The LC of –6 dB appears to be more successful in improving speech intelligibility than the LC of 6 dB. (D. L. Wang, U. Kjems et al, 2009)[59]. (iii) T-F representation for very high resolution, each speaker's parts in a mixture are separate from each other (S. T. Roweis and colleagues 2000) [61]. In these kinds of situations, IBM can almost break a combination down into its parts. A side note: Broadband interferences like noise and echo don't change no matter how many people there are. (iv) Related binary masks have been demonstrated to function well in ASR. Along with missing-data ASR, tools for discovering missing data and other methods for utilizing IBM to improve ASR outcomes have been developed. (v) By Wang et al., IBM noise processing can produce a speech, according to the researchers. During this experiment, it is used to alter the noise's speech-like characteristics (SSN). The spectrum of speech-shaped noise is very similar to that of genuine speech. They found IBM modulated noise to be nearly understandable at low frequencies (e.g., 16 bands).

IBM Estimation Based on Local SNR Estimates

Below given literature talks about calculating SNR in every time-frequency unit, with few examples. These solutions frequently make use of a short-term noise power spectrum estimate. The SNR and, as a result, a T-F mask can be calculated using the predicted noise power. The IBM may be easily determined using the genuine local SNR information, as shown in Equation (2.1).

El-Maliki and Drygajlo (1999) [28] propose the negative energy criterion, which can be used to create masks based on noise estimates. We'll go over some noise-estimation approaches first and then talk about how they can be used to calculate the IBM. When it comes to voice improvement, noise (and SNR) estimation is a common issue, particularly when it comes to spectral subtraction (M. Berouti R et al. 2002)[21]. The frequent misperception is that noise persists throughout a speech and that the first few frames are 'noise-only.' The spectral energy of these frames is averaged to obtain a noise estimate.

Vizinho(1999) [71], Josifovski [35], and Cooke et al.(2001) [25] employ similar estimations. Because of the nonstationarity of noise, such approaches frequently produce incorrect IBM estimates. To estimate noise in nonstationary settings, more complicated algorithms have been proposed. VAD-based methods (A. Korthauer et al. 1999)[72], For instance, Hirsch's histogram-based approaches and recursive algorithms for noise estimation.

Seltzer et al. 2004 [50] approximated the noise similar to Hirsch's approach available in each sub-band, which is then used to estimate masks. Any noise-estimation approach can be effortlessly comprehensive to approximate the SNR at each T-F unit by deriving an approximation from the clean speech power spectrum (M. Berouti et al. and S. Boll et al.1979) [21][23]. By subtracting noise power from noisy spectrum the speech power can be easily calculated. [21][23]. The speech power is determined by subtracting the noise power from the measured noise spectral power. An additional feature is the creation of a spectral floor, below which all estimates are automatically rounded up. There are also a variety of direct SNR estimating methods that have been documented.

Nemer et al. 1999[44] state that the local SNR can be estimated using higher-order speech and noise statistics, which assume a sinusoidal model for band-limit speech and a Gaussian model for noise. Tchorz and Kollmeier [55] proposed a supervised technique for SNR estimation. They used psychoacoustic characteristics and a multilayer perception (MLP)-based classifier to figure out the SNR at each T-F unit.

Loizou 2005[73] provides extensive analyses on these topics for interested readers. If the SNR is approximated using a noise estimate and the LC is set to an appropriate value, the IBM can be calculated using Equation (16.1). Although 0 dB is the most obvious choice, different values have been utilized [25, 49]. Local SNR estimates can be converted to soft (ratio) masks using a sigmoid function, which converts them to a real value in the range [0, 1], allowing them to be understood as probability estimates for further processing. A posteriori SNR, which is defined as the ratio of noisy signal power to noise power expressed in dB (P. Renevey et al. 2001)[48], can also be used to build masks. This eliminates the requirement to calculate the clean speaking power and SNR in the surrounding area. Any a posteriori SNR criterion can be expressed in the same manner as a local SNR criterion.

According to Raj and Stern 2005[46], combining an SNR requirement with a negative energy criterion often leads to higher-quality masks. In practice, noise-estimation algorithms perform well in stationary contexts but poorly in nonstationary ones. Despite this, SNR-based techniques are still popular due to their ease of use.

Speech Enhancement Algorithms

Voice enhancement techniques are intended to reduce noise and reverberation in the input speech stream. Automatic speech recognition, for example, uses this front-end approach in both single-channel and multichannel speech-related domains. Spectral subtraction,

Xu et al. 2014, Zhao et al. 2017, Jin et al. 2009)[75,76,77] discussed Wiener filtering, and Minimum Mean Square Error (MMSE) estimation have all been pursued monaural speech augmentation (Most crucially, these strategies entail the difficult task of estimating noise power in non-stationary sounds. When it comes to voice augmentation, the accuracy of noise estimates for stationary noises is higher than for non-stationary noises. Spectral subtraction, on the other hand, produces good results for non-stationary noises, although it can sometimes produce negative values over the projected speech spectrum. It's interesting to note that model-based strategies are aimed at improving the performance of voice enhancement algorithms in non-stationary noisy situations. Improved speech enhancement methods are based on models like the Hidden Markov model (HMM), Independent Component Analysis (ICA), and Non-negative Matrix Factorization (NMF)(Weninger et al. 2015, Jaureguiberry et al. 2016)[78,79]. LSTM layers are also used in recent speech-enhancement techniques, which use deep learning models with Long Short Term Memory (LSTM) layers.

Kolboek et al. 2016, Weninger et al. (2014)[80,81] introduced speech enhancement with deep LSTM-RNN architecture and successfully integrated it with a speech segregation system to achieve better evaluation measures. Beamforming techniques, BSS-based methods, and neural network-assisted algorithms are more advanced options for multi-channel speech enhancement. Through a process known as beamforming techniques, the enhanced form of the predicted target signal is created by merging several signals that are emitted from diverse spatial locations of an environment. When it comes to speech applications, the three most common beamforming algorithms are delay and sum, minimum variation distortionless response (MVDR) beamformer, and multi-channel Wiener filter (MCWF) (Higuchi et al. 2016, Cauchi et al. 2015)[82,83]. They rely on the computation of the target steering vectors/ spatial covariance matrices as well as the noise spatial covariance matrix in order to function correctly.

Zhang et al. (2017)[84] used deep neural networks and beamforming techniques to create a binaural speech segregation system. Furthermore, a combined method of speech augmentation and speech segregation demonstrates the ability to deal with complicated difficulties in a multisource reverberant setting.

Distance Estimation based on Sound Source

In early 1972, researchers focused their efforts on developing and implementing algorithms for a range of processors, including the Generalized Cross Correlation (GCC), phase transform (PHAT), and Smoothed Coherence Transform (SCOT) (Benesty et al. 2008, Brown et al. 1994)[85,86].

Based on the direct-to-to reverberant ratio, Lu et al. (2010)[98] came up with a good way to figure out where the energy comes from. The author used a reverberation time-dependent binaural equalization cancellation method to show a new way to figure out the direct-to-reverberant ratio for distances greater than 2 m.

DiBiase et al. 2001, Chen et al. 2005, and Frost et al. 1972) [87, 88, and 89] developed models of sound source identification and microphone array signal processing applications using these methodologies. Full source localization, which is computer-assisted, can be used to identify the azimuth angle and distance between the user and a desired target sound source in a noisy and reverberant environment.

Lu et al. 2011, Nguyen et al. 2016 [90,91]. Distance tracking is crucial for a range of acoustic applications, such as intelligent hearing aids, speaker recognition systems, auditory scene analyzers, and audio surveillance systems. The microphone-assisted speaker distance detection system employs pattern recognition and feature extraction approaches (Georganti et al. 2013, Spille et al. 2011, Bishop 2006) [92,93,94].

Georganti et al. 2011, Hioka et al. 2011, Vesa et al. 2009 [95,96,97] tells that it has been found that a number of studies have looked into the effects of room reverberation and volume on distance perception models as well as directivity and outside noise. However, the computational methods for determining the source distance are simpler than those for determining the azimuth localization. Historically, researchers have estimated the distance of a sound source using parameters such as the Interaural Cross-Correlation (ICC) value, pitch coherence, the Direct to Reverberant Ratio (DRR), and energy sequence related to transients (e.g., the center of mass).

Hioka et al. (2011)[96] in his paper explains the method for sound localization. The Direct to Reverberant Ratio was used to develop a spatial correlation matrix model for determining the distance between a sound source and a listener (DRR). When microphone arrays are necessary, this notion has been implemented successfully. The magnitude squared coherence of binaural channels was proposed by Vesa et al. (2009) [107] as a distance perception model. Gaussian maximum-likelihood classification is taught well using the properties learned from these audio channels.

A new study (Georganti et al., 2013)[99] describes an enhanced machine learning-based approach for estimating the distance between two voice sources that are based on the statistical characteristics of the voice source's distance. Additionally, distance estimation models for single- and dual-channel microphone recordings have been successfully constructed using distance-dependent statistical characteristics (Georganti et al. 2011 and 2013)[95,99].

Monaural System

When a spoken signal reaches natural auditory systems, it is frequently accompanied by additional sound sources. However, listeners can hold conversations in a variety of situations. The "cocktail party" effect (E. C. Cherry 1953) [9] is a well-known example of this phenomenon. For computers, it makes sense to allow them to distinguish between the object source and other sources of interference in the same manner as humans do. Automated speech recognition (ASR), speaker identification, audio retrieval, and digital content management are just few of the numerous uses for an effective separation system. As a result, researchers are becoming increasingly interested in voice separation and signal processing in general.

Blind source separation (A. K. Barros et al. 2002)[113] and spatial filtering (H. Krim et al. 1996)[114] are two examples of broad methods for speech separation. These methods necessitate the use of many sensors. However, there is only one sensor in many applications, including telephony and audio recovery; therefore, a monaural solution is expected. Because only one sensor signal might be employed in monaural separation scenarios, it is significantly more difficult and yet an open subject for researchers to investigate. Despite the fact that monaural speech separation remains a difficult task, the human auditory system possesses an extraordinary capacity for it, compelling researchers to continue their investigations into human auditory perception.

In 1990, A. S. Bregman et al. introduced the notion of auditory scene analysis (ASA) for the first time (A. S. Bregman et al. 1990)[24]. According to him, the auditory system may divide acoustic data into streams corresponding to distinct sources using ASA principles. His ASA research suggests a novel approach to the problem of monaural speech separation. As a result, computational auditory scene analysis has generated considerable interest (CASA).

G. J. Brown et al. and various others in [116] discuss for speech separation, numerous CASA systems based on ASA principles have been proposed. It is possible to achieve speech segregation in these systems without making any significant assumptions about the interference's acoustic qualities. There are two fundamental phases in CASA systems: segmentation and grouping (synthesis)[24]. Each sensory segment of audio input should originate from a single source, according to segmentation. Segments that are likely to originate from the same source are grouped together during the grouping stage. CASA research began with an examination of the simplest data-driven technique. Input data such as pitch, onset, offset, AM rate, and so on can be used to derive information about the target speech from this type of CASA system. CASA research has changed from a data-driven to a knowledge-based schema-driven approach over the last decade.

To aid with separation, an increasing amount of higher-level knowledge is being integrated into primitive CASA systems (D. P. W. Ellis et al. 1996, N. Roman et al. 2003, D. Godsmark et al. 1999) [117][119][120]. While recent huge developments in knowledge-based CASA research have resulted in the integration of numerous new forms of knowledge into CASA systems, knowledge about voice perceptual quality has not been paired with them. While most CASA systems are designed to operate in a noisy environment, the signal-to-noise ratio (SNR) is used to assess their overall performance. While the CASA technique improves the SNR of speech following separation and reduces noise, this does not necessarily imply that speech quality is improved in perception. However, it is sometimes assumed that the higher the signal's SNR, the higher the perceived quality of the signal; this is not necessarily the case. Using perceptual quality evaluation systems in conjunction with CASA systems, researchers have developed an approach that enhances

both SNR and perceptual quality of speech separation. While most CASA systems are evaluated based on their signal-to-noise ratio (SNR), the signal-to-noise ratio (SNR) is used to evaluate their overall performance.

Subjective assessment approaches employ listener panels to rate speech quality on a scale of one to five, with one indicating poor speech quality and five indicating great speech quality. The mean opinion score (MOS) is calculated as the average of the listeners' ratings. Although it is the most dependable procedure, it is also the most time and money-consuming, making it unsuitable for frequent or urgent applications. On the other hand, objective measurement approaches that obviate the need for a listener panel in favor of a computing algorithm can overcome these shortcomings. Nonintrusive and intrusive quality evaluation methodologies have been created with the goal of accurately representing subjective ratings of voice signal quality. Invasive measurements, which use some type of distance metric between clean and degraded speech sounds to do so, predict the subjective MOS. Because nonintrusive evaluation is based solely on the test voice signal, estimating objective speech quality is more difficult. While nonintrusive models have been presented in [121], the ITU's objective quality measurement standard method P.563 for nonintrusive was just recently issued.

Hu and Wang's model (2004) [118] Suggested that because speech separation applications lack reference speech signals, an intrusive methodology may be ineffective; thus, the nonintrusive method is advised. As a result, we choose to analyze speech quality using the P.563 technique. After determining the objective quality assessment method, the only remaining task is to establish how to integrate it with a segregation mechanism. They established a link between speech quality and CASA processing based on the features of the fundamental CASA system, specifically With regards to the CASA segmentation results; they applied speech quality evaluation to identify higher-quality segments with fewer interference sources and monitor them using the pitch contour as a separation cue. Finally, we may apply speech quality analysis to determine which segments were excluded from the foreground stream. We can then revisit the initial classification and rearrange the segments to improve the final grouping performance.

Peng Li et al. 2006[8] explains that, in speech and signal processing, monaural speech separation is a difficult problem. Figure 2.3 shows how a monaural speech separation system operates. A single microphone was used to record the sound from sources A and B. A sound or voice separation system can distinguish between A and B sound sources.

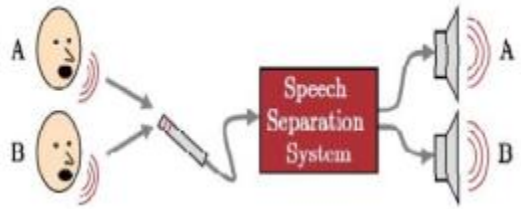

Figure 6.1 Monaural Speech separation system

The cocktail party problem, alternatively referred to as speech segregation, is the difficulty of differentiating object speech from ambient or background noise (Yang Shao et al. 2008, Ming Tu et al. 2014) [7] [2]. Monaural speech segregation involves making monaural recordings using only one microphone and attempting to separate speech and sound. To create a process similar to the human auditory system, recoding using just one microphone may be adequate. Robust speaker and speech identification, audio information retrieval, and hearing aid design are merely some of the real-world uses for this technology [8]. Signal processing, despite decades of work, still has trouble distinguishing monophonic speech. A variety of strategies have been used to overcome the problem of monaural speech segregation.

Peng Li, Yong Guan, Bo Xu, and Wenju Liu's[7], 2006 Computational Auditory Scene Analysis (CASA) and Objective Quality Assessment explored the use of spectral subtraction and Weiner filtering as examples of approaches in their paper "Monaural Speech Separation Based on Computational Auditory Scene Analysis." His presentation focused on computational audio scene analysis (CASA) and objective speech quality evaluation (OSQE) (OQAS).It was proposed here that CASA be used in conjunction with a new method of evaluating the quality of a person's voice (OQAS). With a higher SNR and an average opinion score, the accuracy of speech separation can be improved (MOS). The Hu and Wang model serves as the foundation for the CASA system in the suggested concept (Yuxuan Wang et al. 2013)[4]. For the purpose of distinguishing between resolved and unresolved harmonics, this model is a simple CASA that relies on temporal continuity and cross-channel correlation. Segments are generated and organized using this method based on the regularity of those segments. In addition, a technique known as unresolved harmonics segregation is used to maintain temporal continuity.

As a result, while dealing with spoken discussions, the Hu and Wang model performs nearly as well as many other knowledge-based CASA systems (Yuxuan Wang et al., 2013) [4]. Hu and Wang's model was chosen by the author because it uses the concept of a time-frequency mask, also known as an Ideal Binary Mask (IBM). The audio masking phenomenon inside a crucial band supports the concept of binary masking. In the case of a specific frequency band, a weaker signal is hidden by a stronger one. For human voice intelligibility, the optimal binary mask is particularly successful. It has a great user interface for automatic voice recognition (ASR). Additionally, it simplifies the process of utilizing the CASA system in conjunction with the OQAS algorithm. The author has proved that the proposed strategy enhances the SNRs and the majority of perceptual properties of the split talks significantly. When compared to existing separation or enhancement systems, it was discovered that the proposed method was more effective at processing the monaural speech separation problem than the alternatives in the study.

Yuxuan Wang et al. (2013) [4]. They initially compute a coarse pitch contour utilizing speech split by a dominant pitch to get a precise pitch contour. After that, it's tweaked to fit within psychoacoustic restrictions. Ming Tu et al. (2014) [2] has discussed another use of CASA is voice activity detection (VAD), which is extensively used in ASR, mobile communications for managing discontinuous transmission systems, and a variety of noise tracking techniques for speech augmentation. The accuracy requirements for VAD systems are increasing as the number of voice applications in the real world increases. However, background noise, particularly non-stationary noise, remains an issue for VAD.

The author Ming Tu et al. 2014 [2] explains that feature extraction and decision-making are described as the two fundamental components of a typical VAD system. VAD systems frequently use time-domain characteristics such as energy and zero-crossing rate, as well as spectral-domain characteristics such as spectral difference and DFT coefficients, cepstral-domain characteristics such as Mel-frequency cepstral coefficients (MFCC), and harmonicity-based characteristics such as harmonic structure-based VAD characteristics and DFT harmony. Statistical model-based methods and machine learning-based methods are utilized in decision-making. Computational auditory scene analysis was used to uncover two new VAD features, which the researchers studied in-depth (CASA). There are two basic approaches: one uses the Gaussian Mixture Model, and the other relies on the Variance Analysis of Differences (VAD). Instead of DFT coefficients, GFCC coefficients are extracted from the cochleagramme in the proposed approach, and these features are used to distinguish speech from noise in noisy signals. The GFCC of speech and noise is likewise modeled using the Gaussian mixture model. The proposed approaches' performance is compared to that of many known algorithms. In the job of VAD, the results showed that CASA-based features outperformed various traditional features. The results were assessed using the TIMIT database, and it was determined that GFCC extraction is superior to MFCC extraction. Prior literature has focused on the contrast between spoken and unvoiced speech. Despite the fact that computational auditory scene analysis has been extensively used to recover spoken speech from monoaural mixtures, unvoiced speech separation has gotten relatively little attention. Unvoiced communication is more receptive to influence than voiced communication because of the low intensity and lack of harmonic structure. [13].

Ke Hu 2011[13] explained that using a new method proposed in this, unvoiced speech may be separated from non-voiced speech interference begin, the cross-channel correlation approach proposed here reduces false speech and periodic noise, both of which are undesirable. It is possible to estimate the noise energy of unvoiced intervals by comparing it to the noise energy of neighbouring voiced intervals. To decode silent communication, we must first segment it and then categorize it, as described in the previous section. To create time-frequency segments, unvoiced intervals were segmented using spectral subtraction, which was then used to construct the segments. Using basic thresholding and Bayesian classifiers, unvoiced speech segments are categorized according to the frequency of their unvoiced speech attributes. After rigorous research and comparison, it was discovered that the proposed technique is computationally efficient and considerably enhances the performance of unvoiced speech segregation in unidirectional speech. The article provides an unvoiced speech segmentation algorithm based on a disparately voiced speech that can be applied to a variety of situations. The dormant T-F units of the estimated voiced binary mask were used to estimate noise energy and subtract it from the mixture, resulting in unvoiced segments spaced at regular intervals between voiced segments. Periodic signals will be turned off at that point in time. While the speaker was not speaking, background noise was being analyzed and deleted from the recording system. Calculate an unvoiced interval by averaging the mixing energy of two adjacent voiced intervals with inactive T-F units and dividing the result by the number of active T-F units. T-F analysis can be performed with a 64-channel gammatone filter bank. 64-channel filter banks, rather than a 128-channel filter bank, save computation time by a factor of more than two when compared to the former.

Binaural Systems

Sayers and Cherry's (1995)[126] model was one of the first binaural hearing models, as it connected the lateralization of binaural signals to their interaural cross-correlation. Grouping sources according to their shared source location is a frequently used technique in binaural speech processing for isolating target sounds in difficult conditions. The phrases ITD and IID are frequently used interchangeably to refer to the time discrepancies between two ears (IID). ITD occurs when sound reaches each ear with an unequal delay due to differences in the length of the path between the sound source and the two ears. (Normally, binaural recordings need the use of an artificial head in order to capture substantial IID indications.)

Two microphone recordings are used to differentiate between the target and background sounds, allowing you to hear them more clearly. This is accomplished with two microphone recordings in binaural CASA systems. In most binaural systems, the differences in the signals each ear receives are used to figure out where things are, like azimuth (or microphones). The ITD (interaural tone difference) and IID (interaural tone difference) are the two most important indicators in this regard (IID). ITD is the difference in the time it takes for signals to reach each ear. When ITD happens at a frequency greater than 1.5 kHz, the wavelengths are so small that it is hard to figure out how far apart our ears are. The difference in decibel levels between the two ears is caused by the "shadow" effect that the human head has on it. Because ITD makes low-frequency sound components that surround the listener's head sound different, IID doesn't make them sound different (less than 500 Hz).

Durlarch's equalization-cancellation (EC) model and Jeffress' cross-correlation-based ITD estimate approach both had a significant impact on binaural segregation [27]. There are two stages in trying to distinguish the goal in the EC model. The first step is to equalize the noise levels in the signals originating from both ears. During the cancellation stage, the signals from the two ears are deleted. A cleaner objective is achieved by canceling out the noise that was equalized in the previous stage. Because it is based on the similarity of the two ear signals, the Jeffress model is the most widely used in clinical practise. The inter-trial delay is the amount of time it takes for two patterns of neural activity in both ears to acquire maximum correlation before they become indistinguishable (ITD).

To compute ITD, a normalized cross correlation function, C(t,f,τ)

$$C(t,f,\tau) = \frac{\sum_n xL(tTt-nTn,f)xR(tTt-nTn-\tau Tn,f)}{\sqrt{\sum_n x_L^2(t-Tt-nT,f)}\sqrt{\sum_n x_R^2(tTt-nTn-\tau Tn,f)}}$$

----------2.2

For a time lag of, the preceding equation calculates cross-correlation at frequency channel f and time frame t. Left and right ear responses are denoted by xL and xR, respectively. As with the normalised autocorrelation function, the cross-correlation function peaks at the ITD delay.

IID is calculated as the ratio of the mean power of the signals received by each ear:

$$IID(t,f) = 10log\frac{\sum_n x_L^2\,(tTt-nTn,f)}{\sum_n x_R^2\,(tTt-nTn,f)}$$

10

----------2.3

Roman et al.2003 [49] suggested. This is most likely the first classification-based speech segregation system that has ever been implemented. As the first classification-based speech segregation system, IBM estimates based on classifying ITD and IID estimations. Once a target and interference configuration is determined, they found that the intensities of target and mixture have a smooth and predictable effect on ITD and IID values (The azimuths of the target and the interference are referred to as configuration in this context.). As a result, they were able to use the ITD-IID space to determine the frequency channel-specific distributions of target dominant units and interference dominant units. They use a kernel density estimator to represent the distributions in their system. Based on the observed ITD and IID values and the likelihood that the unit is target dominant or interference dominant, binary judgments are formed at each T-F unit. All three of these metrics have improved when compared to IBM's binary masks. The implementation of ITD-IID distributions is complicated by the fact that they are configuration-dependent.

Palomaki et al.(2003) [45] presented an alternative strategy. TThe azimuths of the target and interference are estimated initially in this technique. The cross-correlation function values at the target and interference azimuths are then compared to determine whether a T-F unit is dominantly target or interference. The precedence effect is replicated by low-pass filtering the envelope response of each channel in their system (R.Y. Litovsky et al.1999) [40]. Conserving instantaneous and suppressing sustained responses, lowers the effect of late echoes in reverberant environments. Palomaki et al. used the above approach to estimate binary masks and obtained good ASR performance in reverberant circumstances.

In 2005, (H Steven et al., 2005; Richard M et al., 1995)[127],[128] provide theories that explain how these cues are employed to lateralize sound sources, among other things. The term "straightness" refers to the

process of weighting ITD contributions to ensure consistency across a variety of frequencies. This was made possible by the fact that authentic sound was produced by point sources that maintained constant ITDs across a broad frequency range. A simple and expedient way for identifying a suitable frequency is to seek for an "unwavering" maximum of the interaural cross-correlation function throughout a certain frequency range. In order to locate spectro-temporal elements that have not been altered by distortion sources like as noise, competing talkers, or reverberation effects, a spectrum-like display has been developed. These approaches have the potential to be effective if the undistorted components are accurately discovered. Using interaural correlation, commonly known as ITD, it has been found that binary or continuous masks may be generated that indicate if a signal's regions are comparable to the source signal's sections.

Similarly, Harding et al. (2006 [30]) provided an approach that assumes the azimuth of the target is already known. The combined distributions of ITD and IID in dominating T-F units can be easily determined using a histogram-based approach. Probabilities for target dominance based on ITD and IID observations are calculated using these distributions. When utilized with a ratio mask, the calculated reverberation probabilities can significantly enhance ASR outputs in reverberant situations.

Woodruff and Wang (2010) [60] recently presented a technique for estimating the IBM that utilizes monaural and binaural signals. Using a monaural CASA approach, they first get simultaneous streams, each of which occupies a continuous time interval. The tandem algorithm, which was previously explained, is used in this process. It is then possible to estimate both the azimuths of the streams and their sequentially grouped concurrent streams using binaural cues.

Chanwoo Kim and his colleagues(2009)[129] used a method called interaural phase difference to figure out binary masks in the frequency domain. This method has made a big difference in how well people can recognize things. Precedence effect: When directional signals from the first wavefront (the direct sound) are given more weight than those from the next wavefronts (the reflected sounds), this is called the "precedence effect."[130]. Many studies show that the precedence effect helps people understand speech and keep track of where they are in reverberant places. The precedence effect is a phenomenon that reduces monaural (Keith D Martin 1997)[131] and binaural (W. Lindemann 1956)[132] echoes. It is possible to raise the onsets or first wavefronts to reduce the effects of reverberation. This can also be performed by suppressing the steady-state components of a signal. The Suppression of Slowly Variable Components and the Falling Edge of the Power Envelope (SSF) algorithm was built using this technique (Chanwoo Kim et al. 2010)[124][133], which significantly improved speech recognition accuracy in reverberant conditions. Numerous additional precedence-based processing algorithms have also demonstrated promising results (e.g., [130,[69]).

Michael L Seltzer et al. 2013, Xue Feng et al. 2014 and Hu et al. [122][123][174]Recent years have seen remarkable advancements in speech recognition systems, owing in large part to the discovery and widespread use of machine learning techniques . On the other hand, noise resistance continues to be a worry.

*Piñero et al. 2017,[15]and Almaadeed et al. 2018,[115] added further that v*oice-based user interfaces for smartphones, smart home devices, cars, and other devices are becoming more common, which makes it important for them to be durable. Improved speech recognition accuracy is still a problem when there aren't any stationary noise sources and other bad things, like reverberation. If you've ever been to a "cocktail party conundrum," you can see how well humans do when there are a lot of different people talking at the same time. Even in the most hostile places, human hearing is very strong. When you understand why our sense of hearing is so strong, you can use auditory processing principles to improve your ability to recognize things in noisy or reverberant situations.

Binaural Processing

Lu, YC (2011) (90) discusses Binaural processing is investigated in a variety of difficult conditions, together with reverberation and interfering talkers. The recording technique, as detailed in the instructions, necessitates the use of two microphones, as seen in Figure 2.4. The recording is done in a reverberant environment, with the two microphones set right in front of the speaker. There is an interfering talker in addition to the two microphones, which is positioned at an angle to the two microphones.

Many of the strategies addressed in this study are based on a knowledge of how humans process audio in both the monaural and binaural senses.

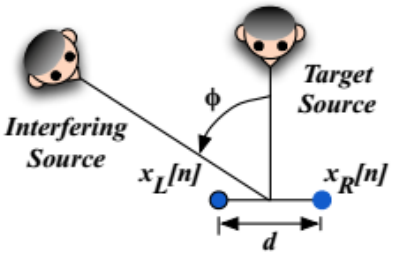

Figure 8.1 2-microphone recording using a target source and an interfering source

Following Figure 2.5 depicts a simplified block diagram of the algorithm discussed in this study. Below are explanations for each of the blocks.

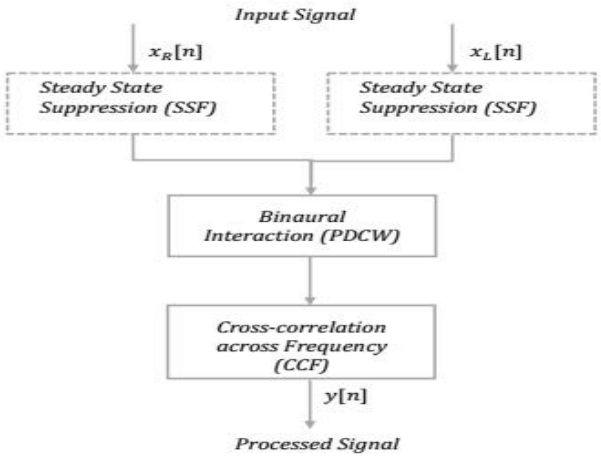

Figure 8.2: Block diagram describing overall algorithm

Steady-state suppression

In the presence of reverberation, steady-state suppression has been found to increase automatic speech recognition accuracy dramatically (ASR). The human auditory system has features like the precedence effect and modulation frequency that led to the use of steady-state suppression in audio processing. The goal of this type of processing is to make more of the input signal sound like direct sound and less like reflected sound.

Chanwoo Kim and Richard et al. (2010) [124,133] explain that this study used the steady-state suppression (SSF) method. Forty gamma frequency channels were originally included in the SSF method. In order to determine the frame-level power of these channels, a low-pass filter is used to analyze and filter the signal. The processed power can be obtained by subtracting the original power contour from a lowpass-filtered representation of the short-term power. The weighting coefficient is calculated by dividing processed power by input. The spectrum weighting factors are then calculated using these weights. Multiplying each spectral weighting factor by the original input signal's spectral spectrum creates the processed signal. This inhibits the power contour's dropping edge, which is particularly useful in reverberant situations for improving ASR performance. We present findings with and without SSF processing in this study. On each microphone channel, steady-state suppression is conducted separately. According to (Richard M Stern et al. 2016) [125], the application of steady-state suppression monaurally is more successful.

Cross-Correlation and across Frequency

Designers offer Cross-Correlation across Frequency (CCF) as a new technique in this study to emphasise input items with consistent frequency distributions. CCF is motivated by the concept of "straightness" weighting, as stated by (R. M. Stern et al. 1998) [151]. This method is intended to smooth a narrow band of frequencies while simultaneously enhancing regions of frequency coherence. CCF processing is depicted in a block diagram in Figure 2.6.

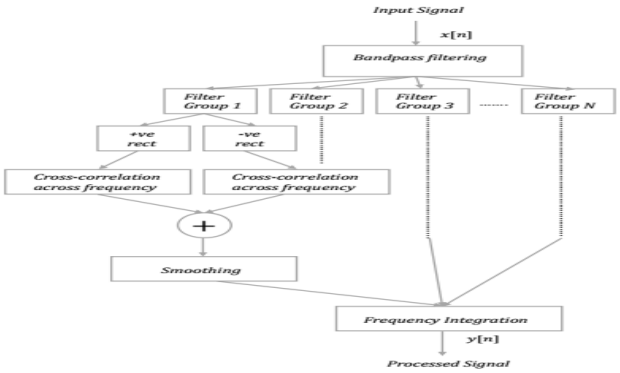

Figure 8.3: CCF algorithm block diagram

This technique closely resembles how the human auditory system processes speech. Bandpass filters are used to simulate the peripheral hearing of the auditory system.[134] The gammatone filters in Slaney's Auditory Toolbox [134] have been modified for our requirements. To compensate for the bandwidth loss caused by squaring the frequency response during the autocorrelation operation, the original gammatone filters' autocorrelation function is computed and adjusted. The ERB scale (Brian CJ 1996) was used to space the filters' center frequencies [135]. Each of these filters is augmented by a satellite filter set. CCF may be conducted over a large frequency range with these satellite filters. Rather than that, N sets of bandpass filters are developed, each of which contains a "central" band and m=2 satellite bands on either side. m is always equal to the number of core bands when expressing the total number of satellite bands.

In Figure 2.6, the N filter groups are denoted as "Filter Group 1," "Filter Group 2," and "Filter Group N." Each of these filter groups consists of a single central band and its associated satellite bands. Each satellite filter in the lth pair has a center frequency determined by the filter group's center band.

$$CB \pm s + \propto^{\frac{m}{2}+1-l}$$

----2.4

CB is an abbreviation for the centre band frequency of a given filter group, and s is a parameter that affects the distance between satellite filters as well as the frequency dispersion on each side of the distance between satellite filters. Compared to when they were at 100 percent, the satellite filters were closer together closer to the centre band and more widely separated further afield when they were at zero. This parameter was determined to be 2500 Hz with N equal to 20 and m equal to 6. The span parameter was determined to be 6 Hz.

The filter outputs for a certain filter group are generated based on the input signal x[n].

$x_{kp}[n]=x[n] * h_{kp}[n]$

-------2.5

Where xkp[n] is the kth band of the pth filter group's filter output, and x[n] is the input. Here, k is a number between 1 and m+1 (m satellite bands plus 1 centre band), while p is a number between 1 and N. A rough model of auditory nerve processing is then applied, which incorporates half-wave rectification of the filter outputs. The filter outputs are negated and half-wave rectified, similar to our earlier work in "polyaural" processing with several microphones (Richard M Stern et al. 2007)[136]. While this part of the processing isn't physiological, it allows for the reconstruction of the full signal, including positive and negative sections. The frequency cross-correlation is then calculated.

Speaker Recognition

Speaker recognition (Zhao et al. 2014, Anguera Miro et al. 2012)[100,101] is the task of recognizing people based on their voice information. The performance of a speaker recognition system is influenced by the combined effects of two factors: noise and reverberation. When noisy test utterances and clean trained utterances are utilized, the performance is lowered. Speech enhancement algorithms, such as spectral subtraction, are incorporated before speaker verification and speaker identification models in a robust speaker recognition system. Few researchers have developed a classifier, such as HMM, that employs speech and sounds separately during training (Dehak et al. 2010 & 2011, Rao et al. 2014)[102,103,104]. The improved final selection is made based on the greatest likelihood at the time of testing. The following algorithms, such as 24 i-vector extraction with PLDA, wiener filtering, and vocal activity detection, are used in state-of-the-art SV systems (Rao et al. 2014, Kanagasundaram et al. 2014 & 2011)[104,105,106]. Some studies have looked at the performance of monaural features such MFCC, GFCC, and amplitude modulation spectrogram under noisy situations to improve speaker recognition resilience (Ming et al. 2007, Zhao et al., 2013; Lei et al. 2012)[107,109,108]. Zhao et al.[109] studied the impact of a noisy reverberant environment on speaker

recognition and performed mask estimation using a deep neural network-based classifier prior to speaker identification. Kanagasundaram et al. [106] proposed many channel compensation approaches to improve the speaker recognition system when the utterance is short. Voice-aided systems, like other biometric modalities, are vulnerable to malicious spoofing attacks, including Automatic Speaker Verification (ASV) systems. Many additional researches have recently concentrated on building classifier fusion-based algorithms to combat spoofing attacks in biometric recognizers (Hanilçi et al., 2018, Sahidullah et al. 2015, Wu et al., 2016)[111, 112,110].

Integrating CASA with ASR

B. Raj et al. (2004)[47] demonstrated the limitations of missing-data ASR when dealing with challenges involving a big vocabulary. They were able to overcome some of the disadvantages associated with using CASA as a preprocessor by using a ratio mask rather than a binary mask and precise mask estimates.

The CASA approaches mentioned in the preceding sections give perceptually inspired strategies for distinguishing the aim from the mixture. Estimating the optimum binary mask has been the main focus. Despite the fact that IBM-based techniques yield good segregation outcomes, merging CASA and ASR has not been as simple as it appears. Using CASA as a preprocessor is a straightforward approach to combine CASA and ASR. The segregated target speech can then be recognized using ASR models that have been trained in clean conditions. This could be a problem. Although the IBM is employed, the resynthesised signal will almost certainly have artefacts that will make identification difficult to achieve. If IBM makes mistakes, this will have a big impact on the performance of these systems. CASA, on the other hand, has been used as a preprocessor in some systems, and it has worked well. Srinivasan et al. [54] proposed one such model. This technique reduces the level of a loud voice by employing a ratio T-F mask. A system called HMM-based ASR, which is trained with Melfrequency cepstral coefficients, recognises the enhanced speech.This system is used to recognise the speech (MFCC). They use Roman et al. [49] to figure out how big the mask is. According to Srinivasan et al.(2010) [7], when the vocabulary size of the recognition job rises, adopting a CASA-based preprocessing strategy may be more advantageous than using missing-data ASR (M. P. Cooke et al. 2001). Hartmann and Fosler-Lussier (2011) [31] made comparisons between ASR systems that use binary masks and those that use information from unmasked T-F units in their recent study on ASR performance and noise reduction.

B. Raj et al. (2004) [47] developed feature reconstruction approaches that increased the noise-resistance of ASR. The recognition process is carried out using an ASR system that is based on HMMs and has been trained in clean environments. In terms of ASR accuracy, reconstructed speech is shown to be much less accurate than IBM-processed speech, although by a few percentage points. Randomly flipping 1s and 0s in the IBM enhances reconstruction only when the number of mask flaws exceeds a predefined threshold. The binary construction of a mask, according to current theory, is anticipated to contaminate the cepstral coefficients (they used PLP cepstral coefficients to build their ASR system). This thing demonstrates the importance of further research into the impact of binary masks on ASR performance. To make use of ASR models trained in clean environments, the methodologies outlined previously modify the characteristics. A way to do this is called "feature compensation" or "source-driven." There are ways to compensate for missing features, like those that use CASA-based algorithms to identify the target [31, 54] and reproduce faulty features. J. Barker et al. 2005, L. Deng et al. 2005, S. Srinivasan et al. 2010[74][56][53] explained that combining CASA and ASR (Another possibility is to change ASR models in such a way that they automatically manage missing or corrupted speech features. These methods are referred to as model compensation or approach to classifier compensation. ASR algorithms for missing data are one type of model compensating mechanism. Additionally, strategies for integrating feature and model correction are available. The approaches developed by Narayanan and Wang (2010)[42] and Karadogan et al. (1958) [36] simplified the process of merging CASA with ASR. They execute ASR on IBMs using a binary pattern classifier. Binary pattern recognition is a significant departure for ASR from established techniques that rely on MFCCs and other fine-grained characteristics of speech. This experiment was inspired by the IBM voice perception study, which established that humans could interpret speech produced by modifying noise. Because noise alone lacks speech information, IBM's binary pattern must be exploited to achieve intelligibility. As a result, the pattern's phonetic information is crucial.

Narayanan and Wang [42] defined it as a system capable of recognizing isolated digits. Convolutional neural networks have proved successful in recognizing handwritten digits and objects (Y. Lecun et al., 1998 [37], [68]). Even if the IBM is computed directly from loud speech using CASA, reasonable results are feasible. As Narayanan and Wang [43] demonstrate, IBMs and traditional speech features such as the MFC complement one another and can be utilized in conjunction to improve their system's overall classification performance. The combined technique can attain accuracy levels comparable to the majority of current phone categorization findings. Additionally, binary pattern features perform well on progressively difficult ASR tasks. In the future, reliable ASR may require characteristics inspired by CASA. The following part will examine an ASR

framework influenced by CASA in greater detail. To improve ASR performance, Srinivasan and Wang's [52] uncertainty transform model uses both feature and model adjustment.

III. Conclusion:

The point that the paper trying to make overview that understanding of ASA involves much more than the understanding of how auditory streams are formed by the alternation of high and low tones in the laboratory. Explanations of ASA – be they in terms of brain processes, computer systems, or the evolution of the nervous system – need to be tested against a wide range of facts about the perceptual organization of sound. And any claim that primitive ASA in non-humans corresponds to primitive ASA in humans also needs to be tested against a wide range of phenomena to see how far the correspondence holds up. Researchers are trying to find out the exact process underlying in the auditory system of human being and making the similar sense in machines. This will open new era for researchers and speech community

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 Elsevier AND Science Direct. 28 March 2008

[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 OnAudio, Speech, And Language Processing, Vol. 14, No. 6, Pp. 2014- 2023 November 2006

[9]. E. C. Cherry, "Some Experiments On Recognition Of Speech, With One And With Two Ears," Journal Of Acoustical Society Of America, Vol. 25, No. 5, Pp. 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]. Taesu Kim, Student Member, IEEE, Hagai T. Attias, Soo Young Lee, Member, IEEE, And Te-Won Lee, Member, IEEE ,"Blind Source Separation Exploiting Higher-Order Frequency Dependencies" IEEE Transactions On Audio, Speech, And Language Processing, Vol. 15, No. 1, January 2007

[12]. Yun Kyung Lee And Oh-Wook Kwon "Application Of Shape Analysis Techniques For Improved Casabased Speech Separation" IEEE Transactions On Consumer Electronics, Vol. 55, No. 1, February 2009

[13]. [13] Ke Hu, Student Member, IEEE, And Deliangwang, Fellow "Unvoiced Speech Segregation From Nonspeech Interference Via CASA And Spectral Subtraction",IEEE Transactions On Audio, Speech, And Language Processing, Vol. 19, No. 6, August 2011

[14]. Guy J. Brown1 And Deliang Wang Separation Of Speech By Computational Auditory Scene Analysis Springer, New York, Pp. 371–402 , 2005

[15]. Piñero, G & Naylor, 'Channel Estimation For Crosstalk Cancellation In Wireless Acoustic Networks', In: Proceedings Of IEEE201International Conference On Acoustics, Speech And Signal Processing (ICASSP), Pp. 586-590, 2017

[16]. [16] Martin Cooke, Guy J. Brown, Malcolm Crawford And Phil Green Endeavour, "Computational Auditory Scene Analysis:Listening To Several Things At Once" , New Series, Volume 17, No. 4, 1993.

[17]. D. Brungart, P. S. Chang, B. D. Simpson, And D. L. Wang, "Isolating The Energetic Component Of Speech-Onspeech Masking With An Ideal Binary Time-Frequency Mask," Journal Of Acoustical Society Of America, Vol. 120, Pp. 4007–4018, 2006.

[18]. E. C. Cherry, On Human Communication. Cambridge, MA: MIT Press, 1957.

[19]. M. P. Cooke, P. Greene, L. Josifovski, And A. Vizinho, "Robust Automatic Speech Recognition With Missing And Uncertain Acoustic Data," Speech Communication, Vol. 34, Pp. 141–177, 2001.

[20]. M. C. Anzalone, L. Calandruccio, K. A. Doherty, And L. H. Carney, "Determination Of The Potential Benefit Of Time-Frequency Gain Manipulation," Ear And Hearing, Vol. 27, No. 5, Pp. 480–492, 2006.

[21]. M. Berouti, R. Schwartz, And R. Makhoul, "Enhancement Of Speech Corrupted By Acoustic Noise," In Proceedings Of The IEEE International Conference On Acoustics, Speech And Signal Processing, 1979.

[22]. P. Boersma And D. Weenink. (2002) Praat: Doing Phonetics By Computer, Version 4.0.26. Available At: Http://Www.Fon.Hum.Uva.Nl/Praat.

[23]. S. Boll, "Suppression Of Acoustic Noise In Speech Using Spectral Subtraction," IEEE Transactions On Acoustics, Speech And Signal Processing, Vol. 27, Pp. 113–120, 1979.

[24]. A. S. Bregman, Auditory Scene Analysis. Cambridge, MA: MIT Press, 1990.

[25]. M. P. Cooke, P. Greene, L. Josifovski, And A. Vizinho, "Robust Automatic Speech Recognition With Missing And Uncertain Acoustic Data," Speech Communication, Vol. 34, Pp. 141–177, 2001.

[26]. L. Deng, J. Droppo, And A. Acero, "Dynamic Compensation Of HMM Variances Using The Feature Enhancement Uncertainty Computed From A Parametric Model Of Speech Distortion," IEEE Transactions On Speech And Audio Processing, Vol. 13, No. 3, Pp. 412–421, 2005.

[27]. [27] N. I. Durlach, "Note On The Equalization And Cancellation Theory Of Binaural Masking Level Differences," Journal Of Acoustical Society Of America, Vol. 32, No. 8, Pp. 1075–1076, 1960.

[28]. M. El-Maliki And A. Drygajlo, "Missing Features Detection And Handling For Robust Speaker Verification," In Proceedings Of Interspeech, Pp. 975–978, 1999

[29]. K. Han And D. L. Wang, "An SVM Based Classification Approach To Speech Separation." In Proccedings Of The IEEE International Conference On Acoustics, Speech And Signal Processing, , Pp. 4632–4635, 2011

[30]. S. Harding, J. Barker, And G. J. Brown, "Mask Estimation For Missing Data Speech Recognition Based On Statistics Of Binaural Interaction," IEEE Transactions On Speech And Audio Processing, Vol. 14, No. 1, Pp. 58–67, 2006

[31]. [31] W. Hartmann And E. Fosler-Lussier, "Investigations Into The Incorporation Of The Ideal Binary Mask In ASR," In Proceedings Of The IEEE International Conference On Acoustics, Speech And Signal Processing, Pp. 4804–4807,2011

[32]. H. Helmholtz, "On The Sensation Of Tone, 2nd Ed. New York: Dover Publishers, 1863.

[33]. G. Hu And D. L. Wang, "Monaural Speech Segregation Based On Pitch Tracking And Amplitude Modulation," IEEE Transactions On Neural Networks, Vol. 15, No. 5, Pp. 1135–1150, 2004.

[34]. G. Hu And D. L. Wang, "Speech Segregation Based On Pitch Tracking And Amplitude Modulation." In Proceedings Of The IEEE Workshop On Applications Of Signal Processing To Audio And Acoustics,Pp. 79–82.2001

[35]. L. Josifovski, M. Cooke, P. Green, And A. Vizihno, "State Based Imputation Of Missing Data For Robust Speech Recognition And Speech Enhancement," In Proceedings Of Interspeech, 1999, P. 2837–2840

[36]. S. G. Karadogan, J. Larsen, M. S. Pedersen, And J. B. Boldt, "Robust Isolated Speech Recognition Using Binary Masks." In Proceedings Of The European Signal Processing Conference, 2010, Pp. 1988–1992.

[37]. Y. Lecun, L. Bottou, Y. Bengio, And P. Haffner, "Gradient-Based Learning Applied To Document Recognition," Proceedings Of The IEEE, Vol. 86, Pp. 2278–2324, 1998.

[38]. N. Li And P. C. Loizou, "Factors Influencing Intelligibility Of Ideal Binary-Masked Speech: Implications For Noise Reduction," Journal Of Acoustical Society Of America, Vol. 123, No. 3, Pp. 1673–1682, 2008.

[39]. Y. Li And D. L. Wang, "On The Optimality Of Ideal Binary Time-Frequency Masks," Speech Communication, Vol. 51, Pp. 230–239, 2009.

[40]. R. Y. Litovsky, H. S. Colburn, W. A. Yost, And S. J. Guzman, "The Precedence Effect," Journal Of Acoustical Society Of America, Vol. 106, Pp. 1633–1654, 1999

[41]. B. C. J. Moore, An Introduction To The Psychology Of Hearing, 5th Ed. London, UK: Academic Press,2003.

[42]. A. Narayanan And D. L. Wang, "Robust Speech Recognition From Binary Masks," Journal Of Acoustical Society Of America, Vol. 128, Pp. EL217–222, 2010.

[43]. A. Narayanan And D. L. Wang, "On The Use Of Ideal Binary Masks To Improve Phone Classification." In Proceedings Of The IEEE International Conference On Acoustics, Speech And Signal Processsing, 2011, Pp. 5212– 5215

[44]. E. Nemer, R. Goubran, And S. Mahmoud, "SNR Estimation Of Speech Signals Using Subbands And Fourth-Order Statistics," IEEE Signal Processing Letters, Vol. 6, No. 7, Pp. 504–512, 1999.

[45]. K. J. Palomaki, G. J. Brown, And D. L. Wang, "A Binaural Processor For Missing Data Speech Recognition In The Presence Of Noise And Small-Room Reverberation," Speech Communication, Vol. 43, Pp. 361–378, 2004.

[46]. B. Raj And R. Stern, "Missing-Feature Approaches In Speech Recognition," IEEE Signal Processing Magazine, Vol. 22, No. 5, Pp. 101–116, 2005.

[47]. B. Raj, M. L. Seltzer, And R. M. Stern, "Reconstruction Of Missing Features For Robust Speech Recognition," Speech Communication, Vol. 43, Pp. 275–296, 2004.

[48]. P. Renevey And A. Drygajlo, "Detection Of Reliable Features For Speech Recognition In Noisy Conditions Using A Statistical Criterion," In Proceedings Of Consistent & Reliable Acoustic Cues For Sound Analysis Workshop, 2001, Pp. 71–74.

[49]. N. Roman, D. L. Wang, And G. J. Brown, "Speech Segregation Based On Sound Localization," Journal Of Acoustical Society Of America, Vol. 114, No. 4, Pp. 2236–2252, 2003.

[50]. M. Seltzer, B. Raj, And R. Stern, "A Bayesian Classifer For Spectrographic Mask Estimation For Missing Feature Speech Recognition," Speech Communication, Vol. 43, No. 4, Pp. 379–393, 2004.

[51]. W. Speith, J. F. Curtis, And J. C. Webseter, "Responding To One Of Two Simultaneous Messages," Journal Of Acoustical Society Of America, Vol. 26, Pp. 391–396, 1954.

[52]. S. Srinivasan And D. L. Wang, "Transforming Binary Uncertainties For Robust Speech Recognition," IEEE Transactions On Audio, Speech And Language Processing, Vol. 15, Pp. 2130–2140, 2007.

[53]. S. Srinivasan And D. L. Wang, "Robust Speech Recognition By Integrating Speech Separation And Hypothesis Testing," Speech Communication, Vol. 52, Pp. 72–81, 2010.

[54]. S. Srinivasan, N. Roman, And D. L. Wang, "Binary And Ratio Time-Frequency Masks For Robust Speech Recognition," Speech Communication, Vol. 48, Pp. 1486–1501, 2006

[55]. J. Tchorz And B. Kollmeier, "SNR Estimation Based On Amplitude Modulation Analysis With Applications To Noise Suppression," IEEE Transactions On Audio, Speech And Signal Processing, Vol. 11, Pp. 184–192, 2003

[56]. D. L. Wang, "On Ideal Binary Masks As The Computational Goal Of Auditory Scene Analysis," In Speech Separation By Humans And Machines, P. Divenyi, Ed. Boston, MA: Kluwer Academic, 2005, Pp. 181–197.

[57]. D. L. Wang, "Time-Frequency Masking For Speech Separation And Its Potential For Hearing Aid Design," Trends In Amplification, Vol. 12, Pp. 332–353, 2008.

[58]. D. L. Wang And G. J. Brown, Eds., Computational Auditory Scene Analysis: Principles, Algorithms, And Applications. Hoboken, NJ: Wiley-IEEE Press, 2006.

[59]. [63] D. L. Wang, U. Kjems, M. S. Pedersen, J. B. Boldt, And T. Lunner, "Speech Intelligibility In Background Noise With Ideal Binary Time-Frequency Masking," Journal Of Acoustical Society Of America, Vol. 125, Pp. 2336–2347, 2009.

[60]. J. Woodruff And D. L. Wang, "Sequential Organization Of Speech In Reverberant Environments By Integrating Monaural Grouping And Binaural Localization," IEEE Transactions On Audio, Speech And Language Processing, Vol. 18, Pp. 1856–1866, 2010.

[61]. S. T. Roweis, "One Microphone Source Separation," In Advances In Neural Information Processing System 13, 2000, Pp. 793–799.

[62]. M. Kawamoto, "Sound Environment Monitoring Method Based On Computational Auditory Scene Analysis", Journal Of Signal And Information Processing, Vol. 8, Pp. 65-77, 2017

[63]. Arkadiy Prodeus, Kateryna Kukharicheva, "Automatic Speech Recognition Performance For Training On Noised Speech", Advanced Information And Communication Technologies (AICT) 2017 2nd International Conference On, Pp. 71-74, 2017.

[64]. Kunkun Songgong , Student Member, IEEE, Huawei Chen , Member, IEEE, And Wenwu Wang , Senior Member, IEEE "Indoor Multi-Speaker Localization Based On Bayesian Nonparametrics In The Circular Harmonic Domain" IEEE/ACM Transactions On Audio, Speech, And Language Processing, Vol. 29, 2021 Pp 1864-1880

[65]. Yuzhou Liu , Student Member, IEEE, And Deliang Wang , Fellow, IEEE "Divide And Conquer:A Deep CASA Approach To Talker-Independent Monaural Speaker Separation" IEEE/ACM Transactions On Audio, Speech, And Language Processing, Vol. 27, No. 12, December 2019 Pp 2092-2102

[66]. Feng Bao And Waleed H. Abdulla , Senior Member, IEEE "A New Ratio Mask Representation For CASA-Based Speech Enhancement" IEEE/ACM Transactions On Audio, Speech, And Language Processing, Vol. 27, No. 1, January 2019 Pp7-19

[67]. Martin Krawczyk-Becker , Member, IEEE, And Timo Gerkmann , Senior Member, IEEE "On Speech Enhancement Under PSD Uncertainty" IEEE/ACM Transactions On Audio, Speech, And Language Processing, Vol. 26, No. 6, June 2018 Pp1144 -1153

[68]. Qiquan Zhang , Aaron Nicolson , Mingjiang Wang , Kuldip K. Paliwal, And Chenxu Wang "Deepmmse: A Deep Learning Approach To MMSE-Based Noise Power Spectral Density Estimation", IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 28, 2020 Pp1404-1415

[69]. Nikolaos Dionelis And Mike Brookes , Member, IEEE "Modulation-Domain Kalman Filtering For Monaural Blind Speech Denoising And Dereverberation", IEEE/ACM Transactions On Audio, Speech, And Language Processing, Vol. 27, No. 4, April 2019 Pp 799-814

[70]. Jaeuk Byun , Student Member, IEEE, And Jong Won Shin , Member, IEEE ,"Monaural Speech Separation Using Speaker Embedding From Preliminary Separation", IEEE/ACM Transactions On Audio, Speech, And Language Processing, Vol. 29, 2021 Pp2753-2763

[71]. A. Vizinho, P. Green, M. Cooke, And L. Josifovski, "Missing Data Theory, Spectral Subtraction And Signal-To-Noise Estimation For Robust ASR: An Integrated Study," In Proceedings Of Interspeech, 1999, Pp. 2407–2410.

[72]. A. Korthauer, "Robust Estimation Of The SNR Of Noisy Speech Signals For The Quality Evaluation Of Speech Databases," In Proceedings Of ROBUST'99 Workshop, 1999, Pp. 123–126.

[73]. P. C. Loizou, Speech Enhancement: Theory And Practice. Boca Raton, Florida: CRC Press, 2007.

[74]. J. Barker, M. P. Cooke, And D. P. W. Ellis, "Decoding Speech In The Presence Of Other Sources," Speech Communication, Vol. 45, Pp. 5–25, 2005.

[75]. Xu, Y, Du, J, Dai, LR & Lee, CH, 'An Experimental Study On Speech Enhancement Based On Deep Neural Networks', IEEE Sig. Proc. Lett.2014, Vol. 21, Pp. 65-68

[76]. Zhao, Y, Wang, ZQ & Wang, DL , 'A Two-Stage Algorithm For Noisy And Reverberant Speech Enhancement', In Proceedings Of ICASSP,2017, Pp. 5580-5584

[77]. Jin, Z & Wang, DL, 'A Supervised Learning Approach To Monaural Segregation Of Reverberant Speech', IEEE Trans. Audio Speech Lang. Process.2009, Vol. 17, No. 4, Pp. 625-638.

[78]. Weninger, F, Erdogan, H, Watanabe, S, Vincent, E, Le Roux, J, Hershey, JR & Schuller, B, 'Speech Enhancement With LSTM Recurrent Neural Networks And Its Application To Noise-Robust ASR' In International Conference On Latent Variable Analysis And Signal Separation Springer,2015 Cham., Pp. 91-99.

[79]. Jaureguiberry, X, Vincent, E & Richard, G, 'Fusion Methods For Speech Enhancement And Audio Source Separation' IEEE/ACM Transactions On Audio, Speech And Language Processing, Vol. 24, No. 7,2016, Pp. 1266-1279

[80]. Kolbœk, M, Tan, Z & Jensen, J, 'Speech Enhancement Using Long Short-Term Memory Based Recurrent Neural Networks For Noise Robust Speaker Verification' In Spoken Language Technology Workshop (SLT), IEEE,2016, Pp. 305-311.

[81]. Weninger, F, Eyben, F & Schuller, B, 'Single-Channel Speech Separation With Memory-Enhanced Recurrent Neural Networks', IEEE International Conference On Acoustics, Speech And Signal Processing (ICASSP),2014, Pp. 3709–3713.

[82]. Higuchi, T, Ito, N, Yoshioka, T & Nakatani, T, 'Robust MVDR Beamforming Using Time-Frequency Masks For Online/Offline ASR In Noise', In IEEE International Conference On Acoustics, Speech And Signal Processing (ICASSP),2016, Pp. 5210-5214.

[83]. Cauchi, B, Kodrasi, I, Rehr, R, Gerlach, S, Jukiü, A, Gerkmann, T & Goetze, S, 'Combination Of MVDR Beamforming And Single-Channel Spectral Processing For Enhancing Noisy And Reverberant Speech', EURASIP Journal On Advances In Signal Processing,2015, Vol. 1, Pp. 61-77

[84]. Zhang, X, Wang, ZQ & Wang, DL, 'A Speech Enhancement Algorithm By Iterating Single- And Multi-Microphone Processing And Its Application To Robust ASR', In Proceedings Of ICASSP, 2017, Pp. 276-280.

[85]. Benesty, J, Chen, J, & Huang, Y, 'Microphone Array Signal Process'. Springerberlin Heidelberg, 2008

[86]. Brown, GJ & Cooke M, 'Computational Auditory Scene Analysis'. Comput. Speech And Language, Vol. 8, 1994,Pp. 297–336

[87]. Dibiase, J, Silverman, H & Brandstein, M , 'Robust Localization In Reverberant Rooms, In Microphone Arrays: Signal Processing Techniques And Applications', Edited By M. Brandstein And D. Ward,Springer,2001, Pp. 157–180.

[88]. Chen, J, Benesty, J & Huang, YA, 'Performance Of GCC- And AMDF-Based Time-Delay Estimation In Practical Reverberant Environments', EURASIP Journal Of Applied Signal Processing,2005, Vol. 1, Pp. 25–36.

[89]. Frost, OL , 'An Algorithm For Linearly Constrained Adaptive Array Processing', In Proc. Of The IEEE, 1972, Vol. 60, No. 8 , Pp. 926-935

[90]. Lu, YC & Cooke, M , 'Motion Strategies For Binaural Localization Of Speech Sources In Azimuth And Distance By Artificial Listeners', Speech Communication, 2011. Vol. 53, No. 5, Pp. 622-642.

[91]. Nguyen, Q & Choi, J, 'Selection Of The Closest Sound Source For Robot Auditory Attention In Multi-Source Scenarios', Journal Of Intelligent & Robotic Systems, 2016,Vol. 83, No. 2, Pp. 239-251.

[92]. Georganti E, May, T, Van De Par, S & Mourjopoulos, J, 'Sound Source Distance Estimation In Rooms Based On Statistical Properties Of Binaural Signals', IEEE Transactions On Audio Speech And Language Processing,2013, Vol. 21, No. 8, Pp. 1727-1741.

[93]. Spille, C, Dietz, M & Hohmann, V, 'Using Binaural Processing For Automatic Speech Recognition In Multi-Talker Scenes', In Proceedings Of IEEE International Conference On Acoustics, Speech,2013

[94]. Bishop, CM 2006, 'Pattern Recognition And Machine Learning (Information Science And Statistics)' Springer. Signal Processing (ICASSP), Pp. 7805 – 7809

[95]. Georganti, E, May, T, Van De Par, S, Harma, A & Mourjopoulos, J, 'Speaker Distance Detection Using A Single Microphone', IEEE Trans. Audio, Speech, Language Process.2011,, Vol.19, No.7, Pp.1949–1961.

[96]. Hioka, Y, Niwa, K, Sakauchi, S, Furuya, K & Haneda, Y, 'Estimating Direct-To-Reverberant Energy Ratio Using D/R Spatial Correlation Matrix Model', IEEE Transactions On Audio Speechlanguage Process, 2011, Vol. 19, No. 8, Pp. 2374–2384.

[97]. Vesa, S, 'Binaural Source Distance Learning In Rooms', IEEE Transactions On Audio Speech Language Process.2009, Vol. 17, No. 8, Pp.1498-1507

[98]. Lu, YC & Cooke, M, 'Binaural Estimation Of Sound Source Distance Via The Direct Reverberant Energy Ratio For Static And Moving Sources', IEEE Transactions On Audio Speech And Language Processing.2010, Vol. 18, No. 7, Pp. 1793-1805.
[99]. Georganti E, May, T, Van De Par, S & Mourjopoulos, J, 'Sound Source Distance Estimation In Rooms Based On Statistical Properties Of Binaural Signals', IEEE Transactions On Audio Speech And Language Processing2013, Vol. 21, No. 8, Pp. 1727-1741.
[100]. Zhao, X, Wang, Y & Wang, DL, 'Robust Speaker Identification In Noisy And Reverberant Conditions', IEEE Trans. Audio Speech Lang Process.2014, Vol. 22, No. 4, Pp. 836 – 845.
[101]. Anguera Miro, X, Bozonnet, S, Evans, N & Fredouille, C, 'Speaker Diarization:A Review Of Recent Research ' , IEEE Trans. Audio Speech Lang. Process.2012, Vol. 20, No. 2, Pp. 356-370.
[102]. Dehak N, Dehak, R, Glass, J, Reynolds, D & Kenny, P, 'Cosine Similarity Scoring Without Score Normalization Techniques' In: Odyssey Speaker And Language Recognition Workshop,2010, P. 15-22.
[103]. Dehak N, Kenny, P, Dehak, R, Dumouchel, P & Oucllet, P, 'Front End Factor Analysis For Speaker Verification', IEEE Transactions Audio Speech Language Process,2011, Vol. 19, No. 4 , Pp. 788-198.
[104]. Rao, KS & Sarkar, S, 'Robust Speaker Recognition In Noisy Environments', Springer International Publishing.2014
[105]. Kanagasundaram, A, Dean, D, Sridharan, S & Vogt, R, 'I-Vector Based Speaker Recognition Using Advanced Channel Compensation Technique', Computer Speech And Language2014,, Vol. 28, No. 1, Pp. 121-140.
[106]. Kanagasundaram, A, Vogt, R, Dean, DB, Sridharan, S & Mason, MW, 'I-Vector Based Speaker Recognition On Short Utterances', In: Proceedings Of The 12th Annual Conference Of The International Speech Communication Association (ISCA),2011, Pp. 2341-2344.
[107]. Ming, J, Hazen, TJ, Glass, JR, & Reynolds, DA'Robust Speaker Recognition In Noisy Conditions', IEEE Transactions On Audio, Speech, And Language Processing,2007, Vol. 15, No. 5, Pp.1711-1723.
[108]. Lei, H, Meyer, BT & Mirghafori, N , 'Spectro-Temporal Gabor Features For Speaker Recognition, In Proceedings Of IEEE International Conference On Acoustics, Speech And Signal Processing (ICASSP),2012, Pp. 4241-4244
[109]. Zhao, X, & Wang, D, 'Analyzing Noise Robustness Of MFCC And GFCC Features In Speaker Identification' In: Proceedings Of IEEE International Conference On Acoustics, Speech, And Signal Processing,2013, Pp. 7204-7208.
[110]. Wu, Z, De Leon, PL, Demiroglu, C, Khodabakhsh, A, King, S, Ling, Z & Yamagishi, J, 'Anti-Spoofing For Text-Independent Speaker Verification: An Initial Database, Comparison Of Countermeasures, And Human Performance', IEEE/ACM Transactions On Audio, Speech And Language Processing (TASLP), 2016, Vol. 24, No. 4, Pp. 768-783.
[111]. Hanilçi, C. 'Linear Prediction Residual Features For Automatic Speaker Verification Anti-Spoofing', Multimedia Tools And Applications, 2018, Vol. 77, No. 13, Pp. 16099-16111.
[112]. Sahidullah, Md, Kinnunen, T & Hanilçi C 2015, 'A Comparison Of Features For Synthetic Speech Detection'. Conference Paper · September 2015 DOI: 10.21437/Interspeech.2015-472
[113]. A. K. Barros, T. Rutkowski, F. Itakura, And N. Ohnishi, "Estimation Of Speech Embedded In A Reverberant And Noisy Environment By Independent Component Analysis And Wavelets," IEEE Trans. Neural Netw., Vol. 13, No. 4, Pp. 888–893, Jul. 2002
[114]. H. Krim And M. Viberg, "Two Decades Of Array Signal Processing Research: The Parametric Approach," IEEE Signal Process. Mag., Vol. 13, No. 4, Pp. 67–94, Jul. 1996.
[115]. Almaadeed, N, Asim, M, Al-Maadeed, S, Bouridane, A & Beghdadi, A, 'Automatic Detection And Classification Of Audio Events For Road Surveillance Applications', Sensors,2018 Vol. 18, No. 6, Pp. 1424-8220.
[116]. G. J. Brown And M. P. Cooke, "Computational Auditory Scene Analysis," Comput. Speech Lang., Vol. 8, Pp. 297–336, 1994.
[117]. D. P. W. Ellis, "Prediction-Driven Computational Auditory Scene Analysis," Ph.D. Dissertation, Dept. Elect. Eng. Comput. Sci., Mass. Inst. Technol., Cambridge, 1996.
[118]. G. N. Hu And D. L. Wang, "Monaural Speech Segregation Based On Pitch Tracking And Amplitude Modulation," IEEE Trans. Neural Netw., Vol. 15, No. 5, Pp. 1135–1150, Sep. 2004.
[119]. N. Roman, D. L. Wang, And G. J. Brown, "Speech Segregation Based On Sound Localization," J. Acoust. Soc. Amer., Vol. 114, Pp. 2236–2252, 2003.
[120]. D. Godsmark And G. J. Brown, "A Blackboard Architecture For Computational Auditory Scene Analysis," Speech Commun., Vol. 27, Pp. 351–366, 1999.
[121]. P. Gray, M. P. Hollier, And R. E. Massara, "Nonintrusive Speech-Quality Assessment Using Vocal-Tract Models," Proc. Inst. Elect. Eng.-Vision, Image Signal Process., Vol. 147, No. 6, Pp. 493–501, Dec. 2000.
[122]. Michael L Seltzer, Dong Yu, And Yongqiang Wang, "An Investigation Of Deep Neural Networks For Noise Robust Speech Recognition," In Acoustics, Speech And Signal Processing (ICASSP), 2013 IEEE International Conference On. IEEE, 2013, Pp. 7398–7402.
[123]. Xue Feng, Yaodong Zhang, And James Glass, "Speech Feature Denoising And Dereverberation Via Deep Autoencoders For Noisy Reverberant Speech Recognition," In Acoustics, Speech And Signal Processing (ICASSP), 2014 IEEE International Conference On. IEEE, 2014, Pp. 1759–1763.
[124]. Chanwoo Kim And Richard M Stern, "Nonlinear Enhancement Of Onset For Robust Speech Recognition.," In INTERSPEECH, 2010, Pp. 2058–2061.
[125]. Richard M Stern, Chanwoo Kim, Amir Moghimi, And Anjali Menon, "Binaural Technology And Automatic Speech Recognition," In International Conference On Acoustics, 2016.
[126]. Bruce Mca Sayers And E Colin Cherry, "Mechanism Of Binaural Fusion In The Hearing Of Speech," The Journal Of The Acoustical Society Of America, Vol. 29, No. 9, Pp. 973–987, 1957.
[127]. Richard M Stern And Constantine Trahiotis, "Models Of Binaural Interaction," Handbook Of Perception And Cognition, Vol. 6, Pp. 347–386, 1995.
[128]. H Steven Colburn And Abhijit Kulkarni, "Models Of Sound Localization," In Sound Source Localization, Pp.272–316. Springer, 2005.
[129]. Chanwoo Kim, Kshitiz Kumar, Bhiksha Raj, And Richard M Stern, "Signal Separation For Robust Speech Recognition Based On Phase Difference Information Obtained In The Frequency Domain.," In INTERSPEECH. Citeseer, 2009, Pp. 2495–2498.
[130]. Patrick M Zurek, "The Precedence Effect," In Directional Hearing, Pp. 85–105. Springer, 1987.
[131]. Keith D Martin, "Echo Suppression In A Computational Model Of The Precedence Effect," In Applications Of Signal Processing To Audio And Acoustics, 1997. 1997 IEEE ASSP Workshop On. IEEE, 1997, Pp. 4–Pp.
[132]. W. Lindemann, "Extension Of A Binaural Crosscorrelation Model By Contralateral Inhibition. I. Simulation Of Lateralization For Stationary Signals," Journal Of The Acoustical Society Of America, Vol. 80, Pp. 1608– 1622, 1986.

[133]. Chanwoo Kim, "Signal Processing For Robust Speech Recognition Motivated By Auditory Processing", Ph.D. Thesis, Carnegie Mellon University, 2010.
[134]. Malcolm Slaney, "Auditory Toolbox Version 2," University Of Purdue, Https://Engineering. Purdue.Edu/˜ Malcolm/Interval/1998-010, 1998.
[135]. Brian CJ Moore And Brian R Glasberg, "A Revision Of Zwicker's Loudness Model," Acta Acusticaunited With Acustica, Vol. 82, No. 2, Pp. 335–345, 1996.
[136]. Richard M Stern, Evandro B Gouvea, And Govindara- Jan Thattai, "Polyaural Array Processing Forautomatic Speech Recognition In Degraded Environments," In Eighth Annual Conference Of Theinternational Speech Communication Association, 2007.