Role of beamforming techniques in the future for IoT, Artificial Intelligence, and Real Time Processing

Sureshkumar Natarajan
Department of Computer and Communication Systems Engineering
Universiti Putra Malaysia
Serdang, Malaysia
suresh45619@gmail.com

Syed Abdul Rahman Al-Haddad
Department of Computer and Communication Systems Engineering
Universiti Putra Malaysia
Serdang, Malaysia
sar@upm.edu.my

Faisul Arif Ahmad
Department of Computer and Communication Systems Engineering
Universiti Putra Malaysia
Serdang, Malaysia
faisul@upm.edu.my

Mohd Khair Hassan
Department of Electrical and Electronic Engineering
Universiti Putra Malaysia
Serdang, Malaysia
khair@upm.edu.my

Raja Kamil
Department of Electrical and Electronic Engineering
Universiti Putra Malaysia
Serdang, Malaysia
kamil@upm.edu.my

Syaril Azrad
Department of Aerospace Engineering
Universiti Putra Malaysia
Serdang, Malaysia
syaril@upm.edu.my

June Francis Macleans
Independent Researcher
Labuan, Malaysia
macleans30june@gmail.com

***Abstract*— Speech recognition from a distance, also known as far-field automatic speech recognition, uses machine learning for processing. However, environmental conditions often corrupt speech recorded from a distance, causing disturbances. To obtain desired speech from corrupted signals, various techniques are used, such as de-reverberation, source separation, denoising, and acoustic beamforming. The aim is to design a robust and multi-condition adaptive system in far-field-based automatic speech recognition systems. This review paper focuses on speech enhancement for the future of speech with progressive technologies like deep learning and machine learning. It highlights the extensive research on beamforming-based speech enhancement over the past few years, based on different techniques, performance, advantages, limitations, and scope for improvement. Finally, this paper explores the smart city applications that benefited from speech enhancement and beamforming.**

Keywords—Speech Enhancement, Beamforming, IoT, Industry 4.0, Smart City, Artificial Intelligence.

I. INTRODUCTION

Speech signal processing, including speech enhancement, speech recognition, and speaker recognition in a system, has evolved to recognize and take the required steps to process the signal for real-time application. Speech signals are complex, but inherently easy to obtain, and there are different noises around. Speech enhancement techniques have become a crucial part of this field as they help reduce noise, improve speech quality, reduce distortion, and minimize unwanted signals to the greatest possible level [1]. This review highlights the extensive research on speech enhancement using beamforming.

Real-time applications such as Google, Alexa, and Siri are prevalent [2]. Researchers are focusing on a far-field environment as applications are prevalent in all interdisciplinary fields of home automation, industry 4.0, healthcare, and "smart city" projects. In the past few years, the field of speech enhancement has grown with the development of denoising techniques such as spectral subtraction [3], Wiener filtering [4], subspace methods [5], and statistical model algorithms [6]. However, the denoising techniques face limitations in suppressing noise in non-stationary environments. Enhancement techniques include the widely used Weighted Prediction Error (WPE) method [7]. WPE uses linear filtering to remove late reverberation by reducing Room Impulse Response (RIR) length, but it does not deal with noise, and performance degrades in unstable RIRs. Another de-reverberation technique is inverse filtering, which uses deconvolution to recover the effects of RIR, although it faces problems in fully implementing the system [8]. Since 2010, the focus on speech enhancement and speech-related systems was shifted to deep learning. Commercial applications of systems that rely solely on speech face intensified background noise; distance involves multiple source images, speech distortion, reverberant speech, and dramatically affects accuracy.

In recent years, the development of speech enhancement techniques has been outlined in several studies [9][10][41]. These techniques have been implemented using neural network (NN) in different domains to extract features, proposed time-synchronized clean and noisy speech pairs, such as feature mapping [11], and time-frequency masking depicted high performance in low SNR at very high reverberant conditions [12].

Joint training with an acoustic model is another effective enhancement with NN that performs significantly better than the conventional method [13]. CLDNN, which is a combination of Convolution Neural Network (CNN), Long short-term memory (LSTM), and deep neural network (DNN), outperforms individual modules by reducing the word error

This research was funded by Research Management Centre, Universiti Putra Malaysia, Grant Putra 9695500.

rate (WER) value. It is interesting to note that CNN with LSTM performed better than DNN with LSTM and the better selection of weight initialization was uniform random weight initialization than Gaussian random weight initialization [14]. Other systems have proposed multichannel speech enhancement for ASR systems by combining acoustic models with deep neural network frameworks [15], and these techniques outperform traditional enhancement beamforming methods. The study reported in [16] claims that the model can be made robust by training the system with various microphone settings to account for data mismatches.

Furthermore, the quality of training provided to the NN is also an essential factor in providing robustness [17]. Training a system with the most realistic test environment can continuously improve the system, but it is practically difficult to obtain such a training set with a very realistic environment. Recently, researchers have generated data for solving issues of far-field or distant talking environments by considering suitable real-time noises and reverberations, such as DIRHA, CHIME-6 challenge, and SRI [18][19]. Based on research so far, this paper focuses on the field of beamforming-based speech enhancement.

II. Contribution of this paper

This paper discusses the recent advancements in speech enhancement using Beamforming and the improvements that have been made in this technology. The article also categorizes the methods used to overcome the main limitations of different techniques. Finally, it outlines the significant changes that have been made to applications using these techniques.

III. Related Work

Beamforming-based speech enhancement can be divided into two main parts: data-dependent (e.g. delay and some beamformer) and data-independent (e.g. minimum variance distortion less response (MVDR), the generalized sidelobe canceller (GSC), and the linear constrained minimum variance (LCMV) beamformer). Hybrid techniques are popular among adaptive beamforming approaches. The study uses the Kronecker product to achieve far-field wideband speech signals using frequency-domain Beamforming of large sensor arrays. This approach splits a Uniform Linear Array (ULA) into smaller virtual versions called virtual ULAs (VULAs) and uses fewer data to estimate the statistics obtained from these VULAs in the form of Minimum Variance Distortion-free Response (MVDR) beamformers from individual arrays. Kronecker products later used combinations of these beamformers resulting in hybrid beamforming techniques, with half used as conventional beamforming and the other half used as MVDR beamformers [20]. When using time-frequency masking, batch and block processing is inefficient as they evolve to frame-by-frame processing in practical applications. MVDR beamforming utilizes frame-by-frame upscaling to boost the signal without delay to eliminate these problems. It combines MVDR with unidirectional Recurrent neural network (RNN) using masking estimation based on the Woodbury matrix identity. This approach successfully outperforms CHIME-3 baseline simulation with a short delay time [21]. With the increasing capabilities and efficient performance of deep neural networks. Acoustic beamforming implements deep eigenvectors as a part of a binary neural network (BNN) which estimates the presence of speech using probability masks obtained from the generalized eigenvalue (GEV) beamformer. It is implemented on CHIME-4 data, enabling better audio quality signals and reducing its computational requirements. In multi-path propagation, there are several challenges like reduction in SNR. A study suggests selecting elements in the codebook of Analog Beam-Formers (ABF) to obtain the highest sum rate. Generally, multi-carrier signals obtain the usage information of a single angle of arrival through learning in the system. The proposed system is a novel machine learning (ML)/DL architecture that continuously operates and avoids spectral efficiency loss due to periodic switching to a dedicated ABF to estimate the required statistics [23]. The CHIME-3 challenge in multi-microphone conditions is a WSJ sentence recorded from a distance using a tablet using six microphones in a bus, street, cafe, and pedestrian noisy condition [24]. The baseline achieves a WER of 33%, but the proposed speech recognition notably achieves a WER rate of 11.4% [25]. However, in 2019, the most effective results were obtained for multi-microphone front-end speech recognition processing at 2.7% WER using an acoustic model topological combination of CNN layers with factorized time-delay neural network (TDNN) layers [26].

The Chime-5/6 challenge examines a dinner party scene with spontaneous dialogue, variable noise environments, reverberation and distortion. Despite the harsh conditions, the baseline can achieve a WER rate of 80%. After better architectural implementation of the backend system, 60% of the WER can be achieved. When source separation is combined with de-reverberation on a multi-microphone system, a WER rate of 43.2% is obtained [26]. CHIME-6 outperforms other systems with a significant 30.5% WER in 2020. Its concepts include deep learning-based iterative speech separation, SNR-based array selection, front-end fusion modeling, and official training data augmentation techniques [27]. The REVERB challenge combines text cues with WSJ datasets. These datasets were recorded from the far field at 2-3 m from the source and microphone array [28]. In a single-channel microphone system, the baseline achieved a WER rate of 44% in 2014. Implementation of the robust system resulted in an improvement from the baseline to 22.2% [29][30].

A significant result was achieved by modifying the front end and adding a multi-microphone system. A WER rate of 6.14% was achieved using an MVDR beamformer implemented to cancel out the direct path signal. The real and imaginary (RI) components of that signal was used to filter the non-target signal for dereverberation [31].

Table I highlights some critical points in several studies involving state-of-the-art speech processing for ordinary and far-field-based speech enhancement. The study focuses on recent years, from 2018 to the present. The widespread use of deep learning, neural network-based mask estimation, and machine learning-based beam selection can address the issue of noisy, reverberant, and complex conditions.

The techniques listed under speech enhancement and beamforming are used to remove noise or reverberation from corrupted speech. Some techniques focus on both, while others aim to suppress noise and reduce speech distortion. However, despite combining these techniques, there has been no overall success in enhancing and eliminating corrupted speech.

Table I. Evaluated study of Beamforming in speech technology

Usage	Advantage	Limitation/Issues
The review discusses the four topics of acoustic impulse response models, spatial filler design criteria, parameter estimation algorithms, and optional post-filtering techniques. [32].	Beamformers with the acoustic impulse response model, the spatial filter design criterion, the parameter estimation algorithm, and optional post-filtering require four transverse perspectives.	However, the application of other technologies requires significant advancements.
Using a binary neural network can estimate the presence of speech with a probability mask obtained from GEV-PAN beamformers corrupted by the four types of noises [22].	Deep eigenvector beamforming obtains better speech quality and is computationally inexpensive.	Accurate estimation of BNN in non-stationary environment is an issue.
MNMF parameters are initiated and incremented to improve performance in unknown noisy environments by using online MVDR beamforming [34].	Even with changing acoustics, the system can adapt to speech from any place in the house.	Computationally complex with inversions of SCM with MNMF. It still depends on steering vector estimations.
The microphone's implementation is external to overcome the possibility of degraded speech in noisy conditions for hearing aid using a beamformer for binaural speech enhancement [35].	Using an external microphone array, the look direction of the hearing aid user controls the beam pattern to improve intelligibility. Also, a binaural enhancement objective measure improves intelligibility.	The analysis of simulated data happens at a shallow SNR level in a reverberant environment. The direction of the head controls the direction of capturing the signal, but it is not necessarily detected in real-time accurately.
To develop versatility in using neural networks on microphone pairs at different spacing and to use time-frequency mask to obtain estimate target and noise covariance matrices used for generalized eigenvalue (GEV) beamforming [36].	Overall experiments improve in SDR from 4.78dB to 7.69dB on various array geometries.	However, considering one interfering source, the latency is around 5s, making it riskier in real time applications.
In order to deal with speech distortions in the presence of intense noise, the proposed system uses mask-based LSTM for noise suppression, and the convolutional encoder-decoder network (CED) for speech restoration uses a spectral mapping technique [37].	In unseen and highly non-stationary environments, achieved results are 0.1 value better than state of the art in terms of PESQ.	However, the improvement is not yet significant, and the method of combination used under higher mismatch may collapse.
For multichannel speech enhancement, Beamforming and post-filtering are combined based on the neural network under the concept of single-channel post filtering with the phase correction [38].	It helps improve intelligibility and speech quality in the presence of multiple speakers due to a combination of post-filtering.	There was an improvement in the PESQ, STOI, and SSNR only after the post-filtering process, but the neural network-based Beamforming did not uplift the results from the MVDR method. Instead, it leads to unnecessary complexities with fewer improvements in values.
A joint parabolic reflector (PR) model is used with a neural beamformer to remove interference speech and background noise from a noisy environment [39].	Under-five different noises, the experiment was carried out at various noise levels. Relative improvements noticed in the noisy conditions were 0.28 in STOI, 1.31 in PESQ, and 11.9 in fwSegSNR.	The system is complex, spectrograms show more significant distortions at high frequencies, and the PR model introduces speech distortions in the target voice.
Kronecker product for far-field broadband speech signals implemented using frequency-domain beamforming of large sensor arrays. It splits the uniform linear arrays into smaller virtual versions called VULAs [40].	Hybrid beamforming techniques use one half as a traditional Beamforming and the other as MVDR, which leads to better extraction of the desired signal.	The drawbacks of each beamforming technique still affect the performance. Moreover, it is yet to find application in a non-stationary environment.
A novel ML/DL architecture for continuously operating the system avoids spectral efficiency losses from periodic switching to a dedicated ABF for the estimation of required statistics [23].	kNN and SVC approaches achieve around 95% of the achievable sum rate with optimal beam selection.	Architecture is complex.
To implement different spatial arrangements using MVDR Beamforming for a hearing aid person in a cocktail party scenario [42].	The microphone position on the forehead is most desirable as it leads to better communication. In addition, adding virtual microphones in the cocktail party scenario increases efficiency in low SNR scenarios.	Due to the input data size of 2s, the delay of the proposed network architecture is too long to be applicable in an actual hearing aid application.
The proposal of U-Net applies to a multi-in and multi-out architecture using neural Beamforming for multichannel speech enhancement [43].	Implementing Skip connection for the convolutional U-Net creates better utility of information.	Linear array formation and higher time consumption are significant limitations for real- time application.
The HRI scenario requires an accurate estimation of the target source location and direction in a time-	In the HRI experiment, the average WER obtained by speech recognition engine is	However, in real-time, any introduced inaccuracies and latency in the

varying acoustic channel [44].	19% lower than publicly available APIs and 34% lower than human testing modalities.	system cannot be afforded.
DNN is combined with a set of AD-HOC microphone arrays to reduce the probability of distant field environment occurrences. In addition, it requires the development of a simple time framework to synchronize channels with different delays [45].	The model with deep learning gave a 2.82 SDR value at high SNR and -6.67 at low SNR.	The novel system faces stability issues, feature extraction, and design under more critical acoustic conditions.
Practically implemented single and multi-microphone systems in video conferencing rooms [46].	It focuses on the real-time scenarios towards far-field multichannel performances of the systems.	This section discusses the limited challenge scope for the experiments.

The architectures utilized in these techniques consider controlled environments up to a maximum of 5 meters. Therefore, the significance of the results is relative.

An improvement in SDR is observed, ranging from 4.78 dB to 7.69 dB, using GEV beamforming on the LibriSpeech ASR dataset. The IEEE Corpus obtains values such as 0.28 in STOI, 1.31 in PESQ and 11.9 in fwSegSNR.

Most of these systems are effective but complex to apply to real-time applications. Time variant-invariant systems are rarely considered, but they are essential when dealing with real-time applications. It is a well-known fact that speech enhancement based on beamforming utilizing AI approaches requires high computational resources and is energy power-consuming. Due to these challenges, implementing real-time applications is difficult [55]. However, some approaches can help overcome these challenges, such as cloud computing services [54]. Recently, a proposed method based on the wireless acoustic sensor network (WASN) platform of distributed microphones has shown real-time performance in speech enhancement based on beamforming using neural networks, smart sensors, and cloud and big data technologies [56][57]. These technologies can help to deliver smart services and applications in real-time.

IV. Internet of Things with Speech Technology

The internet of things (IoT) is an evolving topic that has taken root in every sector of life; the post COVID world has needed rapid advancements in businesses, especially in healthcare. Internet is the backbone of this technology, and due to its ever-growing use, IoT can contribute to the betterment of communication [33, 47]. Speech enhancement based on Beamforming will be essential in applications that will be based on IoT, which is considered the future deployment of most of the technologies. In 2020, Amazon announced hundreds of millions of Alexa users across the globe; this shows the demand and impact value of voice-controlled devices in the future. In addition, these devices use speech signals as commands, quality and intelligibility of these signals is very important to have sufficient experience. Home automation, media entertainment, and security systems are the first movers to the largest customers in the market. Human interaction with machines has rapidly changed from using simple words such as" hi", "thank you", and "sorry" to using speakers, security systems, locks, smart home appliances, thermostats. The expansion of voice shopping is estimated to reach billions by 2022 [48]. Usually, the distance of the speaker from these devices will be more than 4 meters, because of this reason the need of speech enhancement based on beamforming come to the picture and become very urgent. Fig. 1 is a depiction of applications under IoT-based speech intelligence systems integrated with beamforming.

Voice-controlled applications are emerging rapidly; Data hogging is one of the major issues that can be solved. Furthermore, voice data hogging is necessary for the success of voice-controlled networks integrated in real-time for industrial and life science applications [49].

"Last mile" language adoption is the biggest asset in voice-enabled technology that aims to include thousands of global languages from local dialects to the "last mile", where network solutions are hardest to reach. It is an innovative way to preserve the language and cultural history and bring the world closer. The biggest challenge is data availability in all global languages, making it challenging to train artificial intelligence (AI) platforms. In each language, word formation, pronunciation, grammar, and usage are very diverse and vary in complexity from each other.

Privacy by design (PbD) is another requirement for vehicles, homes, stores, workstations, and data security industries. Privacy has become the biggest concern of the future and maintaining personal files of customers has challenges. The new 5G technology has become a cyber security priority, earning customers' trust and protecting products. Privacy by design protects personally identifiable information (PII) in processes and systems [49].

Artificial emotional intelligence is self-explanatory, focusing on more natural expressions and communication with machines. A better understanding of a person's emotional state reveals more about the surroundings and accepting the natural state for better mental health [49].

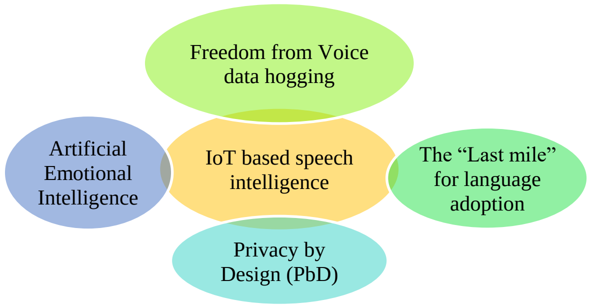
Freedom from Voice data hogging
Artificial Emotional Intelligence
IoT based speech intelligence
The "Last mile" for language adoption
Privacy by Design (PbD)

Fig. 1. Speech based IoT applications integrated with beamforming.

V. Industry 4.0 and Smart City

"Smart city" is a vast concept that includes many merged small-scale techniques. These techniques have a set of required parameters, but the wide range of connecting multiple devices is one essential criterion to cover an entire city base station [50]. Industry 4.0 applies small and large-scale manufacturing unit for better control and monitoring. One of the most important areas that speech enhancement based on beamforming required to be involved with Industry

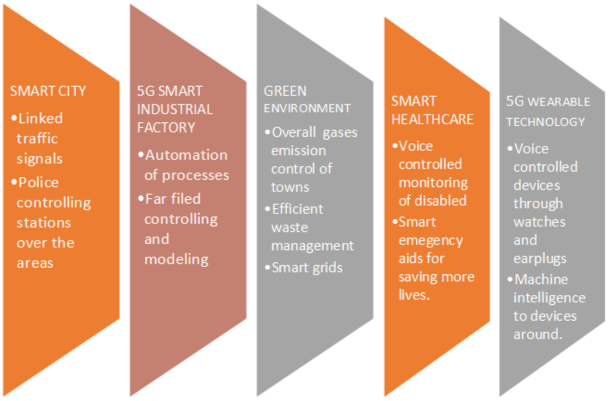
SMART CITY
•Linked traffic signals
•Police controlling stations over the areas
5G SMART INDUSTRIAL FACTORY
•Automation of processes
•Far filed controlling and modeling
GREEN ENVIRONMENT
•Overall gases emission control of towns
•Efficient waste management
•Smart grids
SMART HEALTHCARE
•Voice controlled monitoring of disabled
•Smart emegency aids for saving more lives.
5G WEARABLE TECHNOLOGY
•Voice controlled devices through watches and earplugs
•Machine intelligence to devices around.

Fig. 2. Applications based on beamforming with IoT.

4.0 and smart city applications such as automotive industry. Moreover, car voice controlling is essential service that can help to control the car by the voice command and the quality of these signals is crucial to have sufficient performance [51]. However, in such environment there are many noise sources that make such service very challenging because of the outside noise and interfering passengers signals [52]. Fig. 2 is a description of futuristic applications of Beamforming in different technologies. After the attack of the coronavirus, the world improved healthcare systems. Even before the pandemic, the development of “smart systems” had made great strides, but the need for “smart hospitals” and “smarter technologies” became important and took center stage [53]. For example, the patients with voice pathologies don’t require to refer the doctors and they can easily get the feedback from doctors without the need of leaving their home [54], but in order to have accurate and robust system speech enhancement based on beamforming is required to be integrated with such service.

VI. Conclusion

The creation of revolutionary breakthroughs will be risky. Advances in IoT and machine learning continue to evolve, and the lines of integration have become thinner and thinner over time. Challenges faced are relatively different, but some common issues such as rapid response, real-time dynamic environment simulation, and data security are recurring. This review highlights the role of beamforming in future technologies such as IoT and artificial intelligence. It extensively discusses the application and role of Industry 4.0 concepts. These technologies are undergoing huge improvements as they conquer real-time applications more clearly. The latest research in Table I encourages new researchers to find solutions to complexity, faster speed, better coverage, and better technical sustainability. Hybrid models are popular because of their ability to accumulate individual skills and the new complexity added to them. Exploring ideas, creating balance, and innovating the future based on present challenges is a pertinent summary of this study’s review of speech-based technology.

References

[1] R. Haeb-Umbach, J. Heymann, L. Drude, S. Watanabe, M. Delcroix, and T. Nakatani, “Far-Field Automatic Speech Recognition,” Proceedings of the IEEE, vol. 109, no. 2, pp. 124-148, February 2021.

[2] G. Terzopoulos and M. Satratzemi, “Voice Assistants and Smart Speakers in Everyday Life and in Education,” Informatics in Education, vol. 19, no. 3, 2020, pp. 473–490.

[3] S. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 27, no. 2, pp. 113-120, April 1979.

[4] J. S. Lim, and A. V. Oppenheim, “Enhancement and bandwidth compression of noisy speech,” Proceedings of the IEEE, vol. 67, no. 12, pp. 1586–1604, December 1979.

[5] Y. Ephraim, and Harry L. V. Trees, “A signal subspace approach for speech enhancement,” IEEE Transactions on Speech Audio Processing, vol. 3 no. 4, pp. 251–266, July 1995.

[6] I. Cohen, and S. Gannot, Springer Handbook of Speech Processing. Berlin, Germany: Springer-Verlag, 2008.

[7] Y. Yoshioka, and T. Nakatani, “Generalization of multi-channel linear prediction methods for blind MIMO impulse response shortening,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, no. 10, pp. 2707- 2720, December 2012.

[8] S. T. Neely, and J. B. Allen, “Invertibility of a room impulse response,” Journal of the Acoustical Society of America, vol. 66, no. 1, pp. 165–169, July 1979.

[9] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, “A study on data augmentation of reverberant speech for robust speech recognition,” IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5220-5224, 5-9 March 2017.

[10] B. Li et al., “Acoustic modeling for google home,” Interspeech, pp. 399-403, 20-24 August 2017.

[11] Y. Xu, J. Du, L. R. Dai, and C. H. Lee, “A regression approach to speech enhancement based on deep neural networks,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 23, no. 1, pp. 7-19, January 2015.

[12] A. Narayanan and D. L. Wang, “Ideal ratio mask estimation using deep neural networks for robust speech recognition,” IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 7092-7096, 26-31 May 2013.

[13] Z. Q. Wang, and D. Wang, “A joint training framework for robust automatic speech recognition,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 24, no. 4, pp. 796 - 806, April 2016.

[14] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, “Convolutional, long short-term memory, fully connected deep neural networks,” IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4580-4584, 19-24 April 2015.

[15] T. N. Sainath, R. J. Weiss, K. W. Wilson, A. Narayanan and M. Bacchiani, “Factored spatial and spectral multichannel raw waveform CLDNNs,” IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5075-5079, 20-25 March 2016.

[16] T. N. Sainath et al., “Multichannel signal processing with deep neural networks for automatic speech recognition,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 5, pp. 965-979, May 2017.

[17] M. L. Seltzer, D. Yu, and Y. Wang, “An investigation of deep neural networks for noise robust speech recognition,” IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 7398-7402, 26-31 May 2013.

[18] M. Kumar Nandwana et al., “Robust Speaker Recognition from Distant Speech under Real Reverberant Environments Using Speaker Embeddings,” Interspeech 2018, pp. 1106-1110, 1 September 2018.

[19] Shinji Watanabe et al., “CHiME-6 Challenge: Tackling Multi-speaker Speech Recognition for Unsegmented Recordings,” Proceedings of the 6th International Workshop on Speech Processing in Everyday Environments, pp. 1-7, 4 May 2020.

[20] R. Sharma, I. Cohena, and J. Benesty, “Adaptive and hybrid Kronecker product beamforming for far-field speech signals,” Elsevier, Speech Communication, vol. 120, pp. 42-52, June 2020.

[21] T. Higuchi, K. Kinoshita, N. Ito, S. Karita and T. Nakatani, “Frame-by-frame closed-form update for mask-based adaptive MVDR beamforming,” IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 531-535, 15-20 April 2018.

[22] M. Zohrer, Lukas Pfeifenberger, Gunther Schindler, Holger Froning and Franz Pernkopf, "Resource efficient deep eigenvector beamforming," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3354-3358, 15-20 April 2018.

[23] C. Anton Haro and X. Mestre, "Advanced learning architectures and spatial statistics for beam selection with multi-path," GLOBECOM 2020-2020 IEEE Global Communications Conference, 7-11 Dec. 2020.

[24] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "Multi-microphone speech recognition in everyday environments," Elsevier, Computer Speech and Language, vol. 46, pp.386-387, November 2017.

[25] S. J. Chen, A. S. Subramanian, H. Xu and S. Watanabe, "Building state-of-the-art distant speech recognition using the CHiME-4 challenge with a setup of speech enhancement baseline," Interspeech 2018, pp. 1571-1575, 2-6 September 2018.

[26] C. Zorila, C. Boeddeker, R. Doddipatla and R. Haeb-Umbach, "An investigation into the effectiveness of enhancement in ASR training and test for Chime-5 dinner party transcription," IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pp. 47-53, 14-18 December 2019.

[27] Jun Du et al., "The USTC-NELSLIP systems for CHiME-6 challenge," Proceedings of the 6th International Workshop on Speech Processing in Everyday Environments, pp. 19-23, 04 May 2020.

[28] K. Kinoshita et al., "A summary of the REVERB challenge: State-of-the-art and remaining challenges in reverberant speech processing Research," EURASIP, Journal on Advances in Signal Processing, no. 7, pp. 1-19, 18 January 2016.

[29] M. Delcroix et al., "Strategies for distant speech recognition in reverberant environments," EURASIP, Journal on Advances in Signal Processing, no. 60, pp. 1-15, 19 July 2015.

[30] X. Feng, K. Kumatani, and J. McDonough, "The CMU-MIT REVERB Challenge 2014 system: description and results," REVERB Challenge Workshop, pp. 1-7, 2014.

[31] Z. Q. Wang and D. Wang, "Deep learning-based target cancellation for speech dereverberation," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 941- 950, 28 February 2020.

[32] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A Consolidated Perspective on Multi-Microphone Speech Enhancement and Source Separation," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 4, pp. 692 – 730, 4 January 2017.

[33] S. Markovich-Golan, A. Bertrand, M. Moonen, and S. Gannot, "Optimal distributed minimum variance beamforming approaches for speech enhancement in wireless acoustic sensor networks," Elsevier, Signal Processing, vol. 107, pp. 4-20, February 2015.

[34] K. Shimada, et al., "Unsupervised Speech Enhancement Based on Multichannel NMF-Informed Beamforming for Noise-Robust Automatic Speech Recognition," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 5, pp. 960-971, May 2019.

[35] M. S. Kavalekalam, J. K. Nielsen, M. G. Christensen, and J. B. Boldt, "Hearing aid-controlled beamformer for binaural speech Enhancement using a model-based approach," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 321-325, 12-17 May 2019.

[36] F. Grondin, J. S. Lauzon, J. Vincent, and F. Michaud, "GEV Beamforming Supported by DOA-based Masks Generated on Pairs of Microphones," Interspeech, pp. 3341-3345, 25-29 October 2020.

[37] M. Strake, B. Defraene, K. Fluyt, W. Tirry and T. Fingscheidt, "Speech enhancement by LSTM based noise suppression followed by CNN-based speech restoration," EURASIP Journal on Advances in Signal Processing, no. 49, pp. 1-26, 10 December 2020.

[38] R. Cheng, and C. Bao, "Speech Enhancement Based on Beamforming and Post-Filtering by Combining Phase Information," Interspeech, pp. 4496 – 4500, 25–29 October 2020.

[39] T. Zhang,Y. Geng, J. Sun, C. Jiao and B. Ding, "A Unified Speech Enhancement System Based on Neural Beamforming with Parabolic Reflector," Appl. Sci., vol. 10, no. 7, pp. 1-13, 25 March 2020.

[40] H. Bergaoui, Y. Mlayah, F. Tlilli, and F. Rouissi, "Switching Between Diversity and Spatial Multiplexing in Massive MIMO Systems," UNet 2018: Ubiquitous Networking, vol. 11277, pp. 49-57, 3 November 2018.

[41] X. Cui, Z. Chen, and F. Yin, "Multi-objective based multi-channel speech enhancement with BiLSTM network," Elsevier, Applied Acoustics, vol. 177, pp. 1-13, June 2021.

[42] T. Fischer, M. Caversaccio and W. Wimmer, "Speech signal enhancement in cocktail party scenarios by deep learning based virtual sensing of head-mounted microphones," Elsevier, Hearing Research, vol. 408, pp. 1-12, 1 September 2021.

[43] X. Ren et al. "A Causal U-net based Neural Beamforming Network for Real-Time Multi-Channel Speech Enhancement," Interspeech, pp. 1832-1836, 30 August -3 September 2021.

[44] J. Novoa et al., "Automatic Speech Recognition for Indoor HRI Scenarios," ACM Transactions on Human-Robot Interaction, vol. 10, issue 2, no. 17, pp. 1-30, June 2021.

[45] X-L. Zhang, "Deep ad-hoc beamforming," Elsevier, Computer Speech & Language, vol. 68, no. 101201, pp. 1-18, 31 January 2021.

[46] W. Rao et al., "Conferencingspeech Challenge: Towards Far-Field Multi-Channel Speech Enhancement for Video Conferencing," IEEE Automatic Speech Recognition and Understanding Workshop, 13-17 December. 2021.

[47] H. U. Rehman, M. Asif, and M. Ahmad, "Future Applications and Research Challenges of IOT," International Conference on Information and Communication Technologies (ICICT), pp. 68-74, 30-31 December 2017.

[48] H. Isyanto, A. S. Arifin and M. Suryanegara, "Performance of Smart Personal Assistant Applications Based on Speech Recognition Technology using IoT-based Voice Commands," International Conference on Information and Communication Technology Convergence (ICTC), pp. 640-45, 21-23 October 2020.

[49] J. Y. Lee, and J. W. Lee, "Current Research Trends in IoT Security: A Systematic Mapping Study," Hindawi, pp. 1-25, 13 March 2021.

[50] S. K. Rao and R. Prasad, "Impact of 5G Technologies on Smart City Implementation," Springer, Wireless Personal Communications, vol. 100, no.1, pp. 161-176, 12 March 2018.

[51] M. Vollrath, "Speech and driving–solution or problem?" IET Intelligent Transport Systems, vol. 1, no. 2, pp. 89-94. 2007.

[52] W. Li, Y. Zhou, N. Poh, F. Zhou, and Q. Liao, "Feature denoising using joint sparse representation for in-car speech recognition," IEEE Signal Processing Letters, vol. 20, no. 7, pp. 681-684, 2013.

[53] A. Solanas, C. Patsakis, M. Conti, I. S. Vlachos, V. Ramos, F. Falcone, and A. Martinez-Balleste, "Smart health: A context-aware health paradigm within smart cities," IEEE Communications Magazine, vol. 52 no. 8, pp. 74-81, 2014.

[54] M. S. Hossain, G. Muhammad, and A. Alamri, "Smart healthcare monitoring: a voice pathology detection paradigm for smart cities," Multimedia Systems, vol. 25, no. 5, pp. 565-575, 2019.

[55] K. Tan and D. L. Wang, "A convolutional recurrent neural network for real-time speech enhancement," Interspeech, pp. 3229-3233, 2018.

[56] E. Ceolini, and S. C. Liu, "Combining deep neural networks and beamforming for real-time multi-channel speech enhancement using a wireless acoustic sensor network," IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP), pp. 1-6, October 2019.

[57] M. Chen, Y. Ma, J. Song, C. F. Lai, and B. Hu, "Smart clothing: Connecting human with clouds and big data for sustainable health monitoring," Mobile Networks and Applications, vol. 21, no. 5, pp. 825-845, 2016.