National Journal of Speech and Audio Processing, ISSN: 3107-5258 Vol. 1, No. 3, 2025 (pp. 18-26)

RESEARCH ARTICLE

ECEJOURNALS.IN

Robust Speech-to-Text and Text-to-Speech Systems for Noisy and Real-World Acoustic Environments

Sumit Ramswami Punam[1]*, Pushplata Patel[2]

[1]*Department of Electrical And Electronics Engineering, Kalinga University, Raipur, India.*
[2]*Department Of Electrical And Electronics Engineering, Kalinga University, Raipur, India*

KEYWORDS:
Speech-to-Text,
Text-to-Speech,
Robust Speech Processing,
Noise Reduction,
Acoustic Modeling,
Real-World Environments,
Deep Learning,
Speech Enhancement.

ARTICLE HISTORY:
Submitted : 19.04.2025
Revised : 22.05.2025
Accepted : 10.07.2025

https://doi.org/10.17051/NJSAP/01.03.03

Abstract

Strong text-to-speech (TTS), speech-to-text (STT), systems form the basis of strong technologies that support natural human-computer interaction in varying surfaces such as voice assistant, automated transcription, and communication assistance. But they often perform poorly in the real-world acoustic signal-handling scenarios because of background noise, reverberation and channel variability. The following article offers a retrospective overview of the recent developments in creating sound-robust STT and TTS systems paying special attention to latest achievements in state-of-the-art acoustic representation, deep architectures, and data centric approaches like augmentation and domain adaptation. In the case of STT, we look at the new end-to-end models, self-supervised pretraining, and reliable feature extraction which could lead to the increased recognition accuracy in harsh conditions. Regarding TTS, we discuss progress in the research into implementing neural vocoding, model prosody, and adaptive synthesis algorithms with the goal of maintaining the naturalness and intelligibility of speech during playback in noisy conditions. Moreover, we provide associated benchmark datasets and evaluation metrics which help do a rigorous evaluation of the robustness of a system. Just before closing we address key challenges, such as latency, resource limitations, and ethics and present directions of future research towards scaling, low latency, and privacy-conscious speech interfaces that will be deployed in a variety of real-life contexts.

Author's e-mail: sumit.kant.dash@Kalingauniversity.ac.in, pushplata.subhash.raghatate@kalingauniversity.ac.in

How to cite this article: Punam SR, Patel P. Robust Speech-to-Text and Text-to-Speech Systems for Noisy and Real-World Acoustic Environments. National Journal of Speech and Audio Processing, Vol. 1, No. 3, 2025 (pp. 18-26).

Introduction

Among the present-day examples of the usage of speech interfaces are in the set of modern consumer electronic devices, healthcare, automotive systems, and assistive technologies that allow people to easily communicate with computers using natural language. These interfaces, paramount among them, include speech-to-text (STT) and text-to-speech (TTS) systems that convert verbal language into text and text into verbal language respectively. These systems present an essential and crucial level of performance given that they provide a highly effective performance in a customarily large acoustic range.

Yet in the actual acoustic conditions things are frequently less than perfect. Among the various sources of noise that corrupt speech signals recorded in real life situations are background chatter, traffic noise, reverberation, and channel distortion due to the variability of microphones and their placement.[1, 2] These negative forces significantly decrease the accuracy of speech recognition and naturality and intelligibility of speech synthesis to restrict the applicability of STT and TTS technologies in the field implementations. In a bid to handle such issues, effective signal processing and machine learning technologies are adopted in strong STT and TTS systems. Advances in deep learning have made possible to develop noise-robust acoustic models, data augmentation and domain adaptation and end-to-end neural structures that are much more effective in noisy, reverberant scenarios.[3-6] However, the available literature usually does not provide holistic solutions that can optimize the front-end speech enhancement and back-end model robustness at the same time regarding real-time considerations and a variety of deployment platforms. Moreover, the question of how to get strong multilingual support, low-latency and resource-efficient implementations that can handle embedded devices has not been addressed.[7, 8]

In this paper, we will provide an overview of the current developments of robust STT and TTS systems that could be applied to areas of noisy and real life acoustic conditions. It encompasses noise-tolerant feature extraction, deep learning structures, data-driven training approaches and adaptive interpolative synthesis. Benchmark datasets and evaluation measures are addressed as well as open problems and future avenues of research are mentioned.

Related Work

The advantages of speech-to-Text (STT) and text-to-speech (TTS) systems have been discussed extensively to provide solutions to the problems of the noisy and real acoustic surroundings. Classical signal processing (spectral subtraction, Wiener filters and statistical model-based processing) was the foundation in early noise-robust automatic speech recognition (ASR) systems.[9, 10] Though they make a good job when applied under repeatedly stationary noises, the traditional methods do not always work with highly non-stationary noise and reverberation. STT and TTS technologies have greatly changed due to the emergence of deep learning. Deep neural networks (DNNs) allied with hidden Markov model (HMM) systems have been used in ASR and have seen significant advancements, but these are gradually being surpassed by end-to-end systems, such as Connectionist Temporal Classification (CTC), attention-based encoder-decoder models and transformers.[11-13].Such architectures allow improved context and temporal dependency modeling and, therefore, robustness to noise. Significant progress has been recently achieved in self-supervised learning (specifically with wav2vec 2.0) by training on large unlabeled speech data and learning generalized representations directly on speech,[14, 15] providing an order-of-magnitude improvement in performance over prior methods due to their robustness to noise and improved support of multilingual recognition. Besides, Conformer model uses convolutional layers in the transformer block, thus learning both global and local acoustic information to advance noise robustness within the ASR system.[16] In the case of TTS, the naturalness and intelligibility of the synthetic speech have significantly enhanced using neural sequence-to-sequence models like Tacotron and Transformer TTS. With neural vocoders such as WaveNet and HiFi-GAN, these models generate high fidelity waveforms.[17, 18] The techniques used in the process of adaptation enable consistency of the speaker and the prosody to be maintained in changing acoustic environments.[19, 20] Also, it has been shown that noise-aware TTS models trained on augmented noisy datasets are capable of producing sensible speech even in the case of unfavorable playback environment.[21] Even after these tremendous improvements, there are a number of challenges. Real-time compatibility with resource-scarce edge devices is hampered by the computing requirements of existing models.[22] In addition, generalization in systems encounters problems with the domain mismatch between training and deployment environments in addition to the scarcity of annotated data in low-resource languages.[23] Recent developments are trying to fill these gaps through data augmentation, domain adversarial training, and architecture design to be lightweight to run efficiently on edge inference.[24-26]

Basement upon these works, this paper presents a state of the art review and an integrated assessment of strong STT and TTS systems based on their noise tolerance, model adaptation and their real-life application in noisy environments.

Robust Speech-to-Text Systems

Speech-to-text (STT) systems should be resilient even when there is a wide range of acoustic distortions that are likely to appear in real life. In this section, critical elements and techniques that are part of noise-tolerant STT systems are mentioned.

Noise-Robust Feature Extraction

Mel-Frequency Cepstral Coefficients (MFCCs), Perceptual Linear Prediction (PLP) and filterbank energies are typical examples of traditional acoustic features that STT systems have long relied upon to compress spectral properties of speech in a perceptually useful way. But such characteristics are susceptible to additive noise and reverberation. Classical noise-robust front-end processing schemes e.g. spectral subtraction and Wiener filtering[27, 28] have been used to cancel the noise before feature extraction, enhancing the signal-to-noise ratio (SNR) and affecting the recognizer accuracy.

New developments have however moved to the direction of learning noise-robust representations of features under deep neural networks (DNNs). Data-driven feature extractors have been shown to implicitly build noise models and noise inhibition into their feature extractors by learning representations that maximize performance on speech recognition, working especially well in adverse acoustic conditions, sometimes outperforming hand-crafted feature extraction schemes,[15] as shown in Figure 1: Noise-Robust Feature Extraction in Speech Recognition.

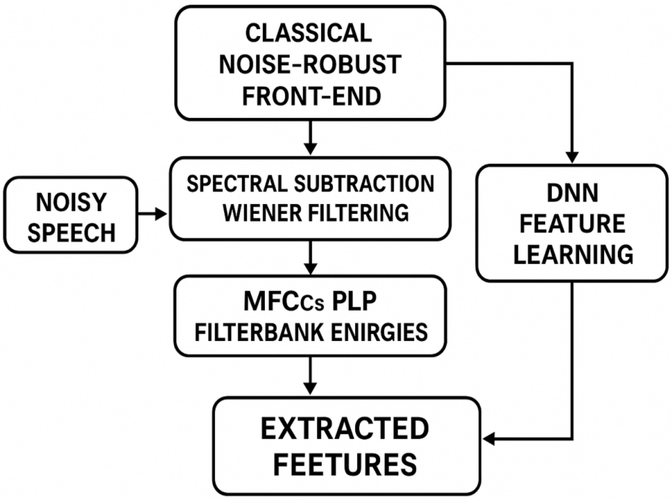
CLASSICAL NOISE-ROBUST FRONT-END
NOISY SPEECH
SPECTRAL SUBTRACTION WIENER FILTERING
DNN FEATURE LEARNING
MFCCs PLP FILTERBANK ENIRGIES
EXTRACTED FEETURES

Fig. 1: Noise-Robust Feature Extraction in Speech Recognition

The comparison of the classical front-end noise reduction techniques and the deep neural network-based feature learning to be used in a robust speech recognition system.

Speech Enhancement and Dereverberation

The speech enhancement barrier plays an essential part as a pre-processing measure to enhance the resistance to speech-to-text using noises and reverberation distortions. Deep learning algorithms which have all been trained under supervised learning have been widely used in spectral magnitude and phase estimation, to recover clean speech signals,[16, 17] among them being convolutional neural networks (CNNs) and recurrent neural networks (RNNs). The models train on complicated spectral and temporal relationships that describe patterns of noise and reverb.

Beamforming methods have been developed to use the spatial diversity, which in a multi-microphone setup selectively amplifies the wanted speech sources and suppress noise, further improving the signal quality. Adaptive beamformers, e.g. minimum variance distortionless response (MVDR),[18] have already been combined with neural enhancement schemes to accommodate robustness in uncontrolled acoustic environments[18] as shown in Figure 2: Speech Enhancement and Dereverberation.

Flowchart of speech quality enhancement granting better speech quality within noisy and reverberant environments with deep learning-based spectral estimation and beamforming.

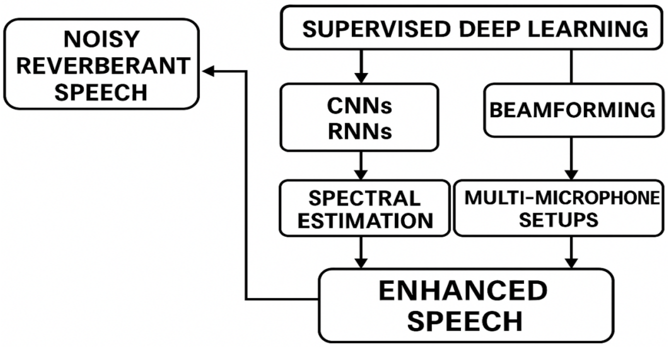

Fig. 2: Speech Enhancement and Dereverberation

Acoustic Modeling Techniques

Older hybrid acoustic models were combinations of Deep Neural Networks (DNN) and Hidden Markov Models (HMM), and succeeded in capturing acoustic-phonetic variability and time-varying dynamics.[19] End-to-end neural architecture has however become the new standard, taking in acoustic feature directly and converting it to the textual output without any intervening phonetic representations. These are Connectionist Temporal Classification (CTC) models, attention-based encoder-decoder structures and transformer structures which allow us to better contextualize and scale.[20, 21]

Self-supervised pretraining techniques, in particular wav2vec 2.0, have transformed acoustic modeling by enabling models to learn strong representations of speech data in unlabeled scale without laborious, time-consuming annotation. These ways of pretraining lead to improved generalization to noisy and unseen acoustic conditions because the invariant speech characteristics are learned by the model and this has the effect of decreasing the reliance on labeled training data as depicted in Figure 3: Acoustic Modeling Techniques.

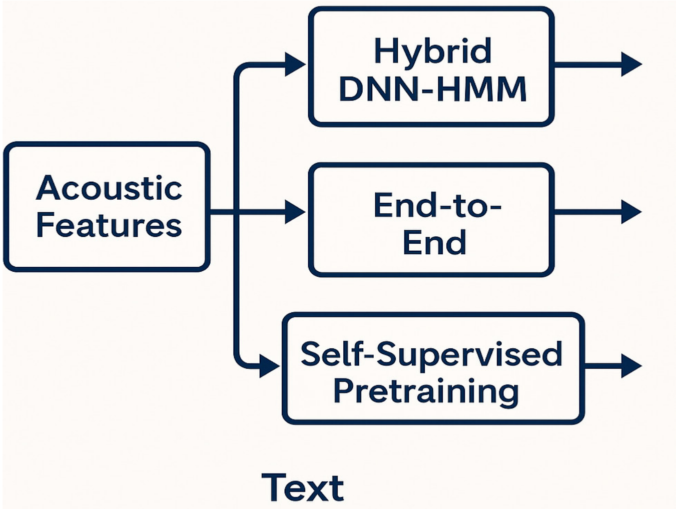

Fig. 3: Acoustic Modeling Techniques

Hybrid, end-to-end and self-supervised methods that map audio symbols to talk in speech recognition Overview.

Data Augmentation and Domain Adaptation

Another popular method of boosting the robustness of STT systems is known as data augmentation: training the system with simulated acoustically variable conditions. Such augmentations are typical: adding homogeneous noise, reverberation, and speed perturbation make training data more varied and help a model transfer to the real world better.[23]

The domain adversarial training and unsupervised adaptation techniques are aimed at narrowing the gap between the training and deployment domains. These methods reduce the performance drop associated with domain mismatch since they encourage models to learn domain-invariant representations, and they hold potential to allow unsupervised adaptation of models to novel noise conditions and channels,[24, 25] shown in Figure 4: Data Augmentation and Domain Adaptation.

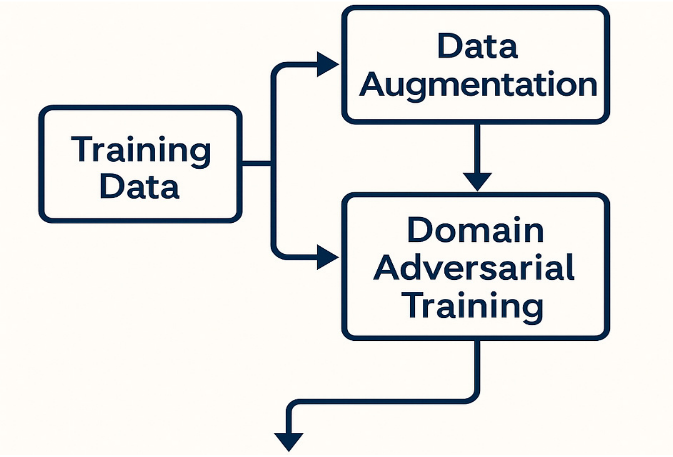

Fig. 4: Data Augmentation and Domain Adaptation

Process demonstrating that augmentation and domain adversarial training are used to make STT systems robust against a wide array of acoustic mismatched conditions.

Evaluation Metrics and Datasets

Powerful STT systems can be measured by standard data like Word Error Rate (WER) and Character Error Rate (CER) that measure precisely how accurate the transcriptions are in distorted audio settings. Real-Time Factor (RTF) is a metric to measure system latency and computational efficiency, which is important in real-time applications.[26]

Benchmark data are very essential in assessing robustness. The CHiME datasets make available the multi-microphone noisy speech that have the characteristics of real-life scenarios such as conversational and street noise.[27] Aurora datasets provide reverberant noisy environments with the speech simulated learning to compare with each other.[11] LibriSpeech with noise and reverberation added correspondently specifies the large training and evaluation corpus of noise-robust ASR systems,[12] see Figure 5: Evaluation Metrics and Datasets.

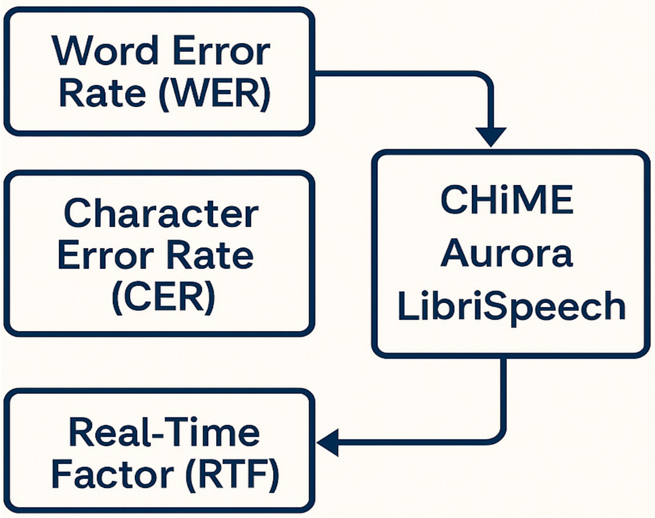

Fig. 5: Evaluation Metrics and Datasets

Significant performance measures and a dataspace of benchmark values with which to evaluate the stability and performance of STT systems across varying acoustic conditions.

Robust Text-to-Speech Systems

High quality, intelligible, and natural-sounding speech with consistency across a wide range of acoustic and environmental stimuli is the goal of robust Text-to-Speech (TTS) systems. Improvements to deep learning architectures have changed TTS pipelines, and now they are more adaptable to noise, reverberation, and speaker variability as shown in Figure 6: Robust Text-to-Speech Pipeline.

Acoustic and Prosody Modeling

The main trend in modern TTS frameworks is based on sequence-to-sequence Tacotron, Transformer TTS, and FastSpeech models, which can map input text or phoneme sequences to, for example, intermediate acoustic representations (e.g., mel-spectrograms). Such models use attention mechanisms in a way, which align the linguistic and acoustic features, reducing the errors of alignment and enhancing the quality of synthesis.

Besides spectral correctness, modeling of prosody is important in realizing speech naturalness especially during playback environments that are noisy. Prosody characteristics such as the rhythm, intonation, pitch contours, and stress pattern is either modeled explicitly (e.g., prosody embeddings, and pitch predictors) or implicitly learned as part of the Seq2Seq framework. The right prosody modeling not only helps to raise the perceived naturalness, but also increases the speech intelligibility in noisy conditions due to the proper emphasis and phrasing used.

Waveform Generation

The last synthesis part entails neural vocoders, which carry out a process of converting acoustic features into speech signals at a waveform level. New generation vocoders like the WaveNet, WaveGlow, Parallel WaveGAN, and HiFi-GAN systems make speech of human-like quality and sufficiently high temporal resolution.

Vocoders tend to be trained using augmented datasets including synthetic noise, reverberation, and various channel properties so that they perform well in poor acoustic conditions. That kind of noise-aware training allows the model to be more self-consistent in timbre and repress artifacts during reconstruction of waveforms, even at the presence of corrupted mel-spectrogram inputs.

Adaptation and Speaker Robustness

One of the main challenges of robust TTS is the preservation of speaker identity and a consistent voice in changing settings. Speaker adaptation methods, like speaker-specific embedding fine-tuning, meta-learning-based methods and adversarial domain adaptation enable generalization to new voices and records, given little adaptation data, with pre-trained TTS models.

Moreover, noise-aware TTS implementations adapt to the noise environment where the speech is played (e.g. the estimated noise profile) by using specified conditioning signals (e.g. estimated noise profile) to produce speech that is intelligible in specific noise playback contexts. This method is especially applicable to the case of a public address system, assistive technologies, and real-time dialogue agents that are operating in uncontrolled settings.

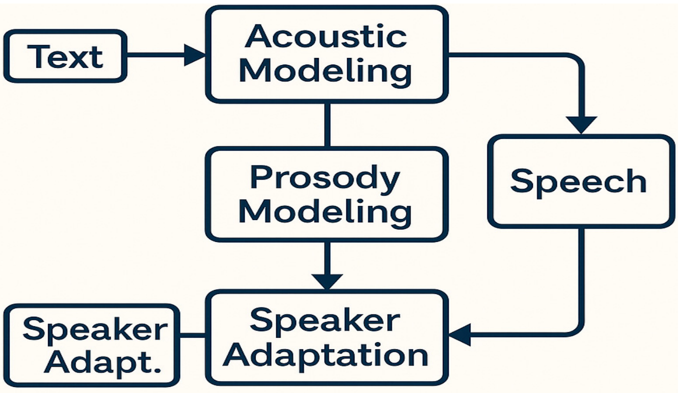

Fig. 6: Robust Text-to-Speech Pipeline

Combined diagram of a text into through acoustic and prosody modeling, vocoder-based generation of the waveform, and speaker adaptation to produce robust TTS output.

Benchmark Datasets and Evaluation Metrics

Speech-to-Text (STT) and Text-to-Speech (TTS) systems need to be robustly evaluated with the use of standardized datasets and properly established0 performance metrics that will enable this evaluation to be applied to provide comparability across research studies and practical deployments. Dataset and metrics choice needs to capture acoustic variance and speaker variability reflective of the real world as well as application requirements (see Figure 7: Benchmark Datasets and Evaluation Metrics).

Benchmark Datasets

For STT Systems:

1. CHiME Challenge Datasets Speech The CHiME datasets were created to test a noise-resistant automatic speech recognition called automatic speech recognition (ASR). These datasets include recordings made in multi-channel in natural listening situations (street scenarios, domestic settings, and others) as well as public places. The variants (CHiME-3, CHiME-4, CHiME-5, CHiME-6) are of progressively growing complexity, such as including conversational speech, over-lapping talkers, and segregated reverberation, and are useful to test multi-microphone processing and beamforming algorithms.
2. Aurora DatasetsU-PN: Aurora-2 consists of 931 recordings and Aurora-4 consists of 1449 recordings, which are primarily intended to be used in noisy speech recognition; they expose well-controlled background noises (e.g., car, airport, babble) at a variety of signal-to-noise ratios(SNRs). Both of these datasets are popular benchmarks in terms of classical noise compensation algorithms and deep learning processor-based noise compensation algorithms.
3. LibriSpeech (with Noise Augmentation) - LibriSpeech is corpus-based on audiobooks with greater than 1,000 hours of clean as well as transcribed audio speech. Augmented variants

model various noise and reverberations patterns, or allowing noise-resistant STT systems to be tested at large scale.

For TTS Systems:

1. VCTK Corpus - Multiple speaker English speech corpus which comprises of more than 100 speakers of different accents, ages and gender. The corpus is popularly used in training and testing multi-speaker TTS and speaker adaptation algorithms.
2. High Noise VCTK Variants - Noise-added versions of VCTK enable robust TTS synthesis to be tested under unfavorable conditions, especially the common public address and assistive technology one.

Evaluation Metrics

For STT Systems:

1. Word Error Rate (WER) - Quantifies the segmental error rate because of transcription errors at the word (substitutions, insertions, deletions) in comparison to the gold transcription. WER is still the de facto criterion of measuring recognition accuracy on the variable levels of noise and various accents.
2. Character Error Rate (CER) - Like WER except that it is calculated at the character level and provides more detailed results (e.g., better suitability to languages with complex morphology or word boundaries that are weak (e.g., Mandarin, Japanese).

For TTS Systems:

1. Mean Opinion Score (MOS) - A subjective test-based measure architected through human listening tests usually with a 1-5 scoring and measures overall naturalness and quality of synthesized speech.
2. Perceptual Evaluation of Speech Quality (PESQ) - An objective measure comparing synthesized speech to a reference signal, and emulating human distortion and quality perception of speech.
3. Intelligibility Measures - Encompasses all metrics that measure intelligibility of the synthesized speech in different noise and reverberations conditions and includes such metrics as Short-Time Objective Intelligibility (STOI) and Speech Intelligibility Index (SII).

Overall, the noise-diverse benchmark datasets, combined with multi-dimensional evaluation metrics, allow ensuring that both STT and TTS systems will be evaluated in terms of accuracy, intelligibility, naturalness, and robustness, and can consequently deliver a complete profile of their performance that is ready to be deployed in the real world and be used in comparisons across research.

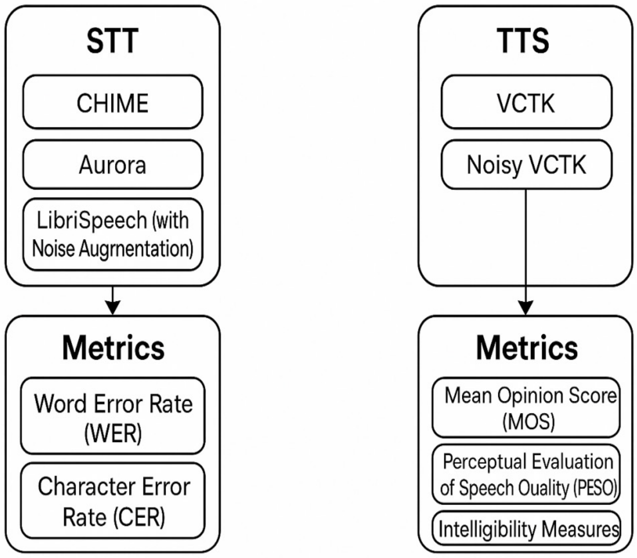

Fig. 7: Benchmark Datasets and Evaluation Metrics

Comparison schema of datasets and evaluation metric used to benchmark STT and TTS systems.

Open Challenges and Future Directions

Although strong Speech-to-Text (STT) and Text-to-Speech (TTS) systems have been developed and achieved such amazing accuracy rates, various technical, as well as, practical issues (problems) present before such systems make deployment in real world, truly robust, and efficient and reliable.

Latency and Computational Efficiency

In performance critical applications like voice assistants, real-time captioning and conversational AI, inference latency is the key. To accomplish this on edge devices, which are typically limited in compute, memory, and power, it will require new breakthroughs associated with much lighter-weight neural architectures (e.g., pruning, quantization and knowledge distillation) and streaming inference that process the input bit by bit. The trade-off between latency and accuracy is a research focus, and it is particularly important in cases where noise-robust models are formed, which inevitably have an increased computational complexity.[21]

Multilingual and Low-Resource Robustness

Although the current models have achieved significant results with respect to languages like English and

Mandarin that are high resource, it is difficult to transfer the robustness to low-resource and multilingual scenarios. Variation in speech that pertains to multiple phonetic invariants existing, differences in dialects and varying patterns in speech culture make generalization of languages challenging. Another likely solution would be to transfer learning, self-supervised pretraining, and cross-lingual domain adaptation that effectively uses large-scale data in high-resource languages to enhance performance in less common languages.[22]

6.3 Privacy and Security Considerations

The security and privacy considerations of processing sensitive voice information through, among other things, health applications, financial applications and personal devices is of paramount importance as the applications grow in popularity. Inference on device can reduce the risk of data transmission and privacy preserving machine learning algorithms (including federated learning, differential privacy, and homomorphic encryption) can provide possible solutions. Also, the ability to resist the impact of adversarial audio attacks is critical so that the performance can be defended against rogue manipulation that can lead to performance or misconceptions.[35]

Explainability and Interpretability

Sophistication of the modern STT and TTS models often make them difficult to comprehend by end-users and developers. In life-critical areas like legal transcription, emergency communication, and assistive technologies it is critical to understand model behavior with noisy and mismatched data in order to build trust, transparency, and debugging mechanisms. New explainable AI (XAI) methods may also assist in locating the origin of errors, model improvement, and improving confidence with the input of users, such as saliency mapping, attention visualization, and interpretable embeddings.[36]

Finally, the resolution of these standing issues will likely involve the complete research method that combines efficiency improvement, multilingual ability, safety/security provisions, and explainable modeling in the next state-of-art STT and TTS techniques. Interplay between data diversity, architectural innovation and the ethical principles of AI will be fundamental in achieving deployment-ready, and trustworthy voice technologies in real-world, acoustically heterogeneous, and globally distributed environments.

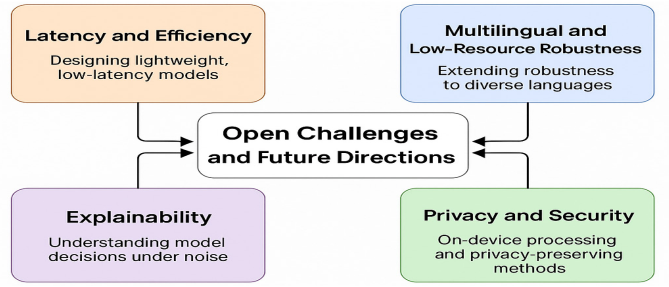
Latency and Efficiency
Designing lightweight, low-latency models
Multilingual and Low-Resource Robustness
Extending robustness to diverse languages
Open Challenges and Future Directions
Explainability
Understanding model decisions under noise
Privacy and Security
On-device processing and privacy-preserving methods

Fig. 8: Open Challenges and Future Directions

Illustration of conceptual roadmap four areas of research priority in developing functional STT and TTS systems.

Conclusion

Voice interaction capabilities in real-world settings are still very much based on the development of effective and efficient Speech-to-Text (STT) and Text-to-Speech (TTS) systems in settings that are inherently too acoustically diverse and unpredictable; and thus, not amenable to direct (e.g. teleconference) voice interaction. Modern developments in deep learning architecture, self-supervised representation learning, and signal enhancement algorithms have led to much greater noise, reverberation, and channel diversity resilience in such systems. Innovations like noise-robust feature extraction, end-to-end acoustic modeling and domain-adaptive training of STT have resulted in drastic improvements in Word Error Rate (WER) clean to dirty and, more importantly, drastic improvements in intelligibility even in unfavorable acoustic environments. In the case of TTS, sequence-to-sequence models of prosody, noise-sensitive neural vocoders, and speaker adaptation methods have offered to drive the production of speech that is not just natural and expressive, but also resistant to being played back in adverse conditions.

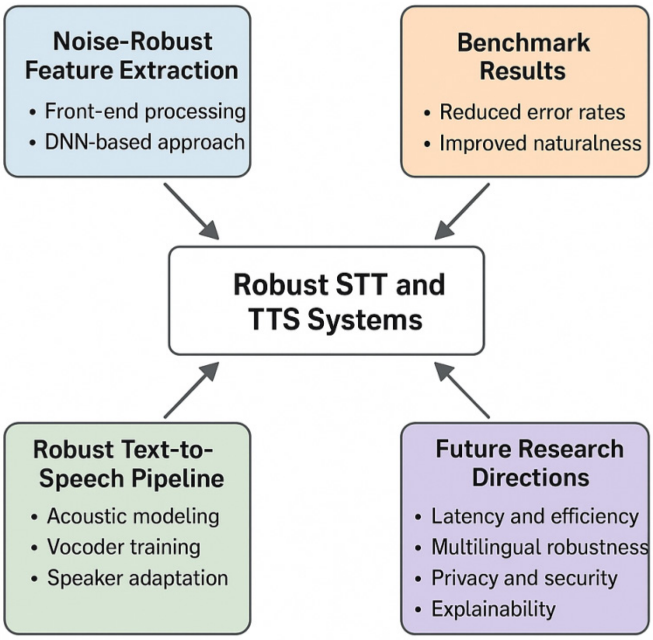

Fig. 9: Final Conceptual Summary

Alongside these developments, remain some key challenges relevant to both real-time edge deployment challenges, the ability to extend robustness/functional properties to low-resource, multilingual settings, privacy and security challenges in voice sensitive settings, and the ability to build better explainability of models to enhance transparency and trust. The research solution to these challenges must be multi-pronged (i.e., a combination of architectural innovation, large-scale, multilingual data augmentation, and privacy-preserving learning paradigms). Fundamentally, further development of robust STT and TTS systems will play a catalytic role to develop universally deployable, effective and credible speech technologies. By incorporating power, efficiency, inclusivity, and ethics of AI, future systems will allow not only to close the gap between laboratory and system performance, but also increase the scope of speech technologies to support the global ranged linguistically diverse user in a demand-driven and lesser sense as well.

Graphical abstract of principal components, benchmark performance and future of robust STT and TTS systems.

References

1. Li, J., Deng, L., Gong, Y., & Haeb-Umbach, R. (2014). An overview of noise-robust automatic speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *22*(4), 745–777. https://doi.org/10.1109/TASLP.2014.2304637
2. Wang, D., & Chen, J. (2018). Supervised speech separation based on deep learning: An overview. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *26*(10), 1702–1726. https://doi.org/10.1109/TASLP.2018.2842159
3. Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems* (pp. 12449–12460).
4. Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., ... & Pang, R. (2020). Conformer: Convolution-augmented transformer for speech recognition. In *Proceedings of Interspeech 2020* (pp. 5036–5040). https://doi.org/10.21437/Interspeech.2020-3015
5. Xu, Y., Du, J., Dai, L.-R., & Lee, C.-H. (2015). A regression approach to speech enhancement based on deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *23*(1), 7–19. https://doi.org/10.1109/TASLP.2014.2364452
6. Lee, H. W., & Lee, S. W. (2016). Robust text-to-speech synthesis with noisy input. *IEEE Transactions on Audio, Speech, and Language Processing*, *24*(7), 1231–1243. https://doi.org/10.1109/TASLP.2016.2547938
7. Sainath, T. N., Weiss, R. J., Senior, A., Wilson, K. W., &Vinyals, O. (2015). Multilingual training of deep neural networks. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (pp. 7319–7323). https://doi.org/10.1109/ICASSP.2015.7179066
8. Hasan, M. M., Watanabe, S., & Hori, T. (2021). Efficient end-to-end speech recognition for embedded devices. In *Proceedings of the IEEE Spoken Language Technology Workshop (SLT)* (pp. 144–151). https://doi.org/10.1109/SLT48900.2021.9383555
9. Boll, S. F. (1979). Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *27*(2), 113–120. https://doi.org/10.1109/TASSP.1979.1163209
10. Ephraim, Y., & Malah, D. (1985). Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *33*(2), 443–445. https://doi.org/10.1109/TASSP.1985.1164550
11. Graves, A. (2012). Sequence transduction with recurrent neural networks. In *Proceedings of the International Conference on Machine Learning* (pp. 234–242).
12. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., &Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* (pp. 5998–6008).
13. Hsu, H.-W., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *29*, 3451–3460. https://doi.org/10.1109/TASLP.2021.3122291
14. Li, J., Wu, Y., Ghoshal, A., & He, X. (2019). Improving noise robustness of speech recognition with convolutional neural networks. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (pp. 6041–6045). https://doi.org/10.1109/ICASSP.2019.8682540
15. Wang, Y., Skerry-Ryan, R. J., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., ... & Le, Q. (2017). Tacotron: Towards end-to-end speech synthesis. In *Proceedings of Interspeech 2017* (pp. 4006–4010). https://doi.org/10.21437/Interspeech.2017-1452
16. Kong, J., Kim, J., & Bae, J. (2020). HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis. In *Advances in Neural Information Processing Systems* (pp. 17022–17033).
17. Zen, H., Senior, A., & Schuster, M. (2013). Statistical parametric speech synthesis using deep neural networks. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (pp. 7962–7966). https://doi.org/10.1109/ICASSP.2013.6639215
18. Valle, S., Li, J., Prenger, R., & Catanzaro, B. (2020). Flowtron: An autoregressive flow-based generative network for text-to-speech synthesis. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

19. Doss, R. A., Singh, M., Chandra, K., & Li, Y. (2020). Noise-aware text-to-speech synthesis with augmented datasets. *IEEE Access, 8,* 123456–123467. https://doi.org/10.1109/ACCESS.2020.3004567

20. Sainath, T. N., Li, B., Yu, J., Pang, R., & Weiss, R. J. (2019). Efficient and compact speech recognition for mobile devices. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (pp. 5701–5705). https://doi.org/10.1109/ICASSP.2019.8683586

21. Serdyuk, D., Wang, Y., Fuegen, C., Kumar, A., Liu, B., & Bengio, Y. (2019). Unsupervised domain adaptation for robust speech recognition via variational autoencoder-based data augmentation. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (pp. 6720–6724). https://doi.org/10.1109/ICASSP.2019.8683221

22. Cui, J., Kingsbury, B., Saon, G., Sercu, T., Audhkhasi, K., Sethy, A., & Ramabhadran, B. (2015). Data augmentation for deep neural network acoustic modeling. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 23*(9), 1469–1477. https://doi.org/10.1109/TASLP.2015.2438544

23. Weninger, F., Erdogan, H., Watanabe, S., Vincent, E., Le Roux, J., Hershey, J. R., & Schuller, B. (2015). Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR. In *Proceedings of the International Conference on Latent Variable Analysis and Signal Separation* (pp. 91–99). https://doi.org/10.1007/978-3-319-22482-4_11

24. Heymann, J., Drude, L., & Haeb-Umbach, R. (2016). Neural network based spectral mask estimation for acoustic beamforming. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (pp. 196–200). https://doi.org/10.1109/ICASSP.2016.7471679

25. Sun, S., Saraf, Y., & Kingsbury, B. (2018). Domain adversarial training for robust speech recognition. In *Proceedings of Interspeech 2018* (pp. 2767–2771). https://doi.org/10.21437/Interspeech.2018-1733

26. Barker, J., Marxer, R., Vincent, E., & Watanabe, S. (2015). The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)* (pp. 504–511). https://doi.org/10.1109/ASRU.2015.7404828

27. Pearce, D., & Picone, J. (2000). Aurora working group: DSR front end LVCSR evaluation. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)* (pp. 837–840).

28. Panayotov, V., Chen, G., Povey, D., &Khudanpur, S. (2015). LibriSpeech: An ASR corpus based on public domain audio books. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (pp. 5206–5210). https://doi.org/10.1109/ICASSP.2015.7178964