Review of Multidisciplinary

https://mdresearchreview.online/index.php/4/about

RM

Advancements in Speech Synthesis and Recognition Technologies through Machine Learning and Phonetics

Aziz Ullah
PhD Scholar, Department of University of Peshawar
Saad Khalid
MPhil Scholar, Department of Political Science University of Peshawar

ABSTRACT

This research explores advancements in speech synthesis and recognition technologies, focusing on the integration of machine learning and phonetics. As voice-based interfaces become increasingly prevalent in various applications, understanding the interplay between phonetic principles and machine learning algorithms is essential for enhancing the accuracy and naturalness of speech technologies. This study employs a comprehensive analysis of recent developments in deep learning techniques, including neural networks and reinforcement learning, to investigate their impact on synthesizing human-like speech and improving recognition systems. Key components such as prosody, intonation, and articulation are examined to identify how phonetic features can be effectively modeled and incorporated into these systems. The findings indicate that leveraging phonetic insights alongside machine learning frameworks significantly enhances the performance of both speech synthesis and recognition, resulting in more intuitive and responsive user interactions. Additionally, the research discusses the implications of these advancements for diverse fields, including accessibility, telecommunications, and artificial intelligence. Ultimately, this study contributes to the understanding of how interdisciplinary approaches can drive innovation in speech technologies, paving the way for more sophisticated and human-centric voice interfaces.

Keywords:**Keywords:** speech synthesis, speech recognition, machine learning, phonetics, deep learning, neural networks, prosody, artificial intelligence.

1. Introduction

Effective Communication is an essential element in human life and is also a field that has increasingly attracted the attention of researchers, especially in the fast increase in speech technology area. The continuous dialogue between developments has resulted in a thoroughgoing integration of many machine learning and phonetics disciplines with speech technology, while the advances in this area have in turn returned on the related areas. This paper aims to provide a broad account of these multi-level trends on some selected topics rather than being an in-depth examination of single subjects. Therefore, considering the broadness of the area, while the theoretical aspects are to be dwelled upon, the technical implementation is also kept in the foreground, and a greater emphasis is laid on the slowing down effects of some sectional developments to their related areas. (Karagthala & Shah, 2022)

Following a brief introduction, a general outline will be given in the subsequent section. Part 2 deals with fundamental researches oriented towards the phonetics and phonology disciplines, that nest the basis of modern-day speech technology applications (Jain & Rastogi, 2019). It is followed by Part 3, dedicated to the basic principles of digital signal processing (DSP) and machine learning (ML), the twin supporting axes of almost all of today's speech technology researches and applications. Some of the most deterring challenges today's speech technology faces are described in Part 4. At this stage, an excursion is taken in trans-disciplinary studies, orienting from speech technology to phonetics disciplines for the understanding of the phonetic properties of acoustically modified speech. Engineering applications are spotlighted in Parts 5 through 13. Finally, the leading trends in today's speech technology researches are outlined, let alone waiving to cover all variety. Way beyond this, it is hoped that this work will be an informational guide suitable for the beginner, as well as for the sectors concerned, and also give a credible account of the most current state of the arts. (Mehrish et al.2022)

2. Fundamentals of Phonetics and Speech Synthesis

Phonetic science comprises three subprocesses: articulatory, acoustic, and auditory phonetics. Articulatory phonetics is concerned with the production or articulation of sounds. During the

production of speech, a set of movements of the vocal tract, principally by the articulators, produces speech sounds. The vibrating air in the vocal tract sets the air outside in motion, causing lung-generated sound waves to travel through the air as sound vibrations. These sound vibrations are then captured at some distance by a listener's ear, and in the inner ear through complex processes they eventually excite a signal which travels to brain centers where these signals are encoded (Tan et al., 2021). One aspect of this production and reception is the main focus of analysis: how to simulate this phenomenon that we perceive as speech. Reinforcement of the production error would appear vital, thus reducing the processing resources needed by listeners. On the production side, error correction through auditory feedback is a known factor in the individuation of human speech, possibly supporting efficient communication (Tits et al., 2019). Humans are remarkable communicators; this phonetic framework highlights some of the foundations that encode their speech so robust. Despite the continuous disturbances, speakers can still produce a largely automatic and highly specialized act that can be understood. Roughly, less than 1% word error rates happen in idealized conversation. Based on phonetic knowledge, various speech synthesis methods have been developed. In a broader sense, synthesizing potentials and properties of any linguistic units can be regarded as text-to-speech conversion. Statistically, the first text-to-speech conversion system was performed by Kempelen in the 18th century, who constructed a talking machine that could produce five short Hungarian fixed sentences. However, he died before finishing his invention. In the history of computer-based speech synthesis, it started during the 1960s with three basic techniques: articulatory synthesis, formant synthesis, and concatenative synthesis. With the release of the first line speech synthesizers in the 1970s, the initial progress in analysis and modeling of human speech was noticed. The early speech synthesizer development that took place in several laboratories worldwide explored a range of different approaches. Major steps forward were made in speech synthesis upgrading the quality of the resultant speech as illustrated by the improvement in the intelligibility of synthetic dialogue. While there were plans that sophisticated high-quality synthesis could be achieved by the mid-1980s, the understanding of the processes and the rules shaping human speech and the fostering of the processing power necessary for such a system were not forthcoming (Brahmi et al.2022). Progress in speech synthesis, particularly synthesis by

rule, was slow. After the hidden Markov model reached scientific maturity in the late 1970s, leading to the modern era of speech recognition, it began to be used as a basis for research into innovative techniques in other areas of speech technology, including speech synthesis. Six TTS methods, including formant, articulatory, and concatenative synthesis, were thus developed. Linguists have understood the relationship between phonetics and speech synthesis and are educated in each other. With the recent advances in machine learning, including deep learning, researchers have endeavored to apply knowledge of machine learning from speech recognition to improve and design speech synthesis methods. Noticeably, SPSS, SLT, WPSLML, ARSW/T, DSCLM, and WCSNT are among the several acronyms used in grammar. However, this study mainly concentrates on reviewing neural speech synthesis. Although a similar term of "neural TTS" already exists, SLT researchers name it in a broader sense. (Reflinda et al.2022)

2.1. Basic Concepts in Phonetics

Phonetics is the scientific field that studies how speech sounds are produced (articulatory phonetics), how they are perceived (auditory phonetics), and their physical properties (acoustic phonetics). This first step in the understanding of speech forms is a crucial element for the solution to speech recognition, synthesis or translation tasks. Frequently, the task is formulated as the conversion of an input speech signal into a sequence of phonemes or into an alternative symbolic representation like pronunciation. Consonants and vowels are the two basic classes into which speech sounds are classified. Consonants are sounds that are produced with a significant constriction somewhere in the vocal tract. Vowels, on the other hand are produced with a relatively open vocal tract. A vowel is characterized by 3 characteristics: the position of the tongue in the mouth (front vs back, high vs low, closed vs open), the position of the lips (rounded or not) and the tenseness of the tongue. Suprasegmentals are prosodic features, such as intonation and rhythm, that span sounds and extend over time. Intonation concerns the pitch or melody of speech, how it rises and falls or goes up and down. Rhythm includes the relative durations of speech sound and their patterns of alternation in speech. Hence, it is important to understand the relative durations of different speech sounds of the same language. Syllables are the rhythmic unit in speech and are easy to perceive once one is aware of the limitations on possible syllables

at the outset of lexical acquisition (Tits et al., 2019). The physical aspects of a sound can be defined in terms of frequency, amplitude, etc. of a sound wave. However, these physical aspects of sound interact in a great many complex ways with human hearing. One of the central topics of acoustic phonetics is to study these complex interactions critically to learn about how the initial physical forms of sounds are heard as particular phonetic qualities. This section describes the basic physical principles according to which speech sounds can be analyzed by using devices that measure the sound signal. Speech samples are represented using appropriate notation systems; the most widely used is the International Phonetic Alphabet (IPA), a standardized phonetic transcription system that represents most (or all) of the speech sounds used in the world's languages, including consonants and vowels (Al-Fraihat et al.2022).

2.2. Speech Synthesis Techniques

Speech synthesis techniques have evolved significantly in response to advances in technology and the associated research in acoustics and phonetics. Voice synthesis engines have emerged as computational and algorithmic tools that automate the generation of speech sound. Rule-based synthesis was the first successful voice synthesis system and reproduces speech through a series of overdetermined rules governing phonetic, prosodic, and co-articulation aspects to generate an artificial signal by concatenating diphones together – this is referred to as "artificial diphthong speech synthesis" (Tits et al., 2019). As technology progressed, these diphone databases were replaced by triphone ones. It is shown that the diphone speech synthesis was modified to produce a unified tri-diphone system using the same set of rules.

Concatenative synthesis is an approach to generating speech waveforms that can create more naturalistic speech, which consists of splicing together pre-recorded segments of speech; this method is considered as the closing part of the analysis by synthesis. However, with the increase in the processing speed of computer platforms, concatenative speech synthesis was replaced by more complex and flexible methods (Tan et al., 2021). Producing speech in this manner has the benefit that all of the components of the natural acoustic signal are present in the diphone inventory upon which the synthesis is based. Speech studies have used two techniques of concatenative synthesis for the production of spoken stimuli in an intelligibility experiment. It

was used to demonstrate that electronically generated stimuli can replicate the relevant phonetic features of naturally spoken input to within an acceptable degree of approximation. Further, as technology advances, so have other methods for the synthesis of speech instead of the traditional phonetic-based rules. This early work in conjunction with early work in automatic speech recognition set the groundwork for the current machine learning-based speech synthesis methods that belong to the state-of-the-art today (Kumar et al., 2022).

3. Machine Learning in Speech Recognition

Machine learning has become an essential element to boost the performance of speech recognition systems, a vital part of voice technologies. Such systems can automatically analyze and interpret spoken language using machine learning algorithms. They can model the complexity of phonetics, linguistics, acoustics, and psychoacoustics of speech signals more effectively as compared to traditional rule-based algorithms. Thereby, the accuracy of the speech recognition system is improved. The increased use of machine learning is observed in various applications, including business transactions, healthcare, security, telecommunications, and candidature. Recent studies and work in such areas has been discussed more on the utility of machine learning algorithms in the field of speech processing as a more efficient and accurate way for spoken language understanding (Jain & Rastogi, 2019).

Speech recognition systems, widely employed in industries, need implementations of extensive machine learning techniques. These learning technologies enable the recognition of patterns in audio signals and hence help in transforming them into texts. This process of transforming audio signals to text includes a couple of more fundamental methodologies like feature extraction, and classification. Feature extraction is the first and most imperative stage. It extracts meaningful information from the raw representation of the signals. This extraction technique creates an intermediate representation of the audio signal, which provides more information about the signal but is comprehensible. This same process is investigated by the brain of any living system. The created intermediate representation is further fed to the classification methodology to transform these intermediate structured audio signals into more comprehensible text (Latif et al., 2020). There are considerable contributions of learning technologies in the classification stage.

Broadly two types of learning are employed in the recognition system, supervised and unsupervised learning. Supervised learning is a kind of recognition system where the system is trained with some labeled data. Labeled data is the data that includes the input and the corresponding output. There are a number of supervised learning models, a neural network, Hidden Markov Model (HMM), Support Vector Machine (SVM), K-Nearest Neighbor (KNN), etc., employed in the training process of the recognition system. Unsupervised learning is used to train the recognition system with the unlabeled data. There are two kinds of learning algorithms in the unsupervised learning method, clustering, and association methods. The principle is to extract knowledge from the unstructured, raw data, patterns observed in the raw data. Since labeled data is very difficult, time-consuming, and expensive to collect as compared to the unstructured data, hence unsupervised learning might be the promising solution to set a speech recognition task without requiring an extensive dataset. Conversely, these models are computationally very intractable and hard to train. Furthermore, there are not always good assignments for the number of clusters (Iskarous & Pouplier, 2022).

3.1. Supervised Learning Algorithms

Supervised learning is the process of training models with labeled datasets to recognize patterns and make predictions based on the input features. Supervised algorithms are designed to learn input-output mapping with a set of examples that consist of input-output pairs (Jain & Rastogi, 2019). An essential part of the modeling process consists of providing labeled data for the algorithms to predict the output accurately. Voice recognition is one of the burgeoning technologies that are currently being employed in many products and services like voice-controlled devices, automatic transcription services, and speech to text applications as well. Several algorithms have been used with varying degrees of success to recognize patterns in speech data including Hidden Markov Models and support vector machines, to name but a few. It is well documented that voice recognition system's accuracy will be higher when it is trained on a labeled dataset. Many researchers have shown results of speaker-dependent isolated word recognition rates between 4% and 5% for systems trained on dictation reading speech. A substantial improvement in recognition rates has been observed when the amount and quality of

the training data used to train the acoustic models is improved. In general, increasing the size of the labeled training data corpus represents an effective method of improving performance in automatic speech recognition systems. It has been successfully demonstrated that the performance of the Voice Activity Detector and thus the completeness of the datasets significantly impact phone recognition rates. Since first described, it has been well understood that supervised learning algorithms have several problems while building models and that many of them are evolved around the task of optimization. A common problem is overfitting and additional post-training model algorithms that fine-tune to a test set or are used as a method of domain adaptation. One of the most popular methods of domain adaptation is to train the model to a large annotated dataset, which is used to create either more annotations, lexical resources, and modeling data or synthetic modeling data (Bhangale and Mohanaprasad2021).

3.2. Unsupervised Learning Techniques

One interesting research area is how unsupervised learning techniques or closely related representations can be used in natural language processing specifically in the context of a word-based straight end-to-end text-to-speech framework. This could include research on linguistics, a few proposed original small unsupervised machine learning techniques, and possible results on various unsupervised-combined representations of words that can be quantitatively compared in the downstream TTS performance. Large volumes of unlabelled data without manual input annotation provide a new opportunity for exploring unsupervised learning and representation techniques for TTS. There are a few unsupervised learning methods that do not require human annotation of the input data. This includes clustering and dimensionality reduction. Clustering can be used to organize similar data into a finite number of groups. This is used in conjunction with methods that compare and group data points based on the similarity of their features. In the context of speech, the features can be the filterbanks' energies in multiple frames and the group of similar data looks for the same phoneme element.

After training, deep clustering assigns each audio to a corresponding speaker cluster. Forced alignment labels are combined using a one-hot vector, reshaped and processed as a mask to force all frames of the same phoneme to a location. On the other hand, dimensionality reduction learns

a simplified or compact representation of the data. Its goal is to reduce the number of input components while preserving the important information in the data. Feature extraction, projection pursuit, regression methods, or manifold learning are popular methods. In automatic speech recognition, a feature extractor is trained as unsupervised learning first using the spectral distance measure for speech recognition. After that, it converts the high-dimensional input observation to a lower-dimensional representation and includes useful information for the recognition task. The main idea of the proposed speech recognition is to use a deep clustering-based unsupervised pre-trained speech feature to improve phoneme segmentation. This segmentation is used to train the classifier for the given speech recognition task or system to align speech signal for speech recognition. The resulting phoneme posterior is used as a phoneme-specific score of confidence.

4. Deep Learning Models for Speech Synthesis

Deep learning models have made notable achievements in the realm of speech synthesis. Works on deep learning are substantial, and the application on the audio synthesis area is engaging. With inputs of audio spectrogram, deep learning models are increasingly capable of generating fluent and contextual audio output. At present, typical deep learning models for audio synthesis involve RNNs, CNNs, and transformers. Besides, it demonstrates how these models improve the performance of speech synthesis through numerous experimental results that have been revealed by related studies. This work aims to provide recent advances in speech synthesis technologies. Furthermore, it creates an DSP-based deep learning text-to-speech (TTS) end-to-end model for aspiring users to use.

RNNs are a class of neural networks with connections between neurons forming a sequence. In simple terms, an RNN processes input sequences across time steps and output the corresponding data. Because the internal structure of the hidden layer neurons contains a memory copy of the past network input information, RNN can process data sequentially. When extended to colossal data sets, RNN models can exploit their ability to learn dependencies which result in excellent tasks of sequential data (Tan et al., 2021). CNNs are also well suited for speech recognition as they are hierarchical, thus learning representations from large and input data. They are beneficial

for the automatic extraction of spectral features for this task. By considering local dependencies and with the help of convolution filters, CNNs model data dependencies across time more effectively than feedforward networks while also addressing the need for input data of fixed dimensions, which has made them more suitable for sequence-to-sequence architecture than feedforward models (Yao et al., 2022).

4.1. Recurrent Neural Networks (RNNs)

Recurrent neural networks (RNNs) are an essential class of architectures in deep learning mostly applied in the context of sequence processing tasks, which include important subfields like speech and language. Neural networks can analyze and learn from massive data sets quickly, but traditional architectures like feedforward and convolutional neural networks do not process their input sequentially. RNNs address this limitation by feeding information back into the network, resulting in hidden states that store a memory of past inputs. The feedback loop processes each element in a sequence and a fixed-size vector of hidden states becomes a representation of the past until that point in time. This makes RNNs extremely expressive models for sequences, and they have seen great use in language modeling (Nagesh Shewalkar, 2018). RNNs applied to the softmax function predict probabilities over time of what should come next in the sequence. Thus trained on a sufficiently large collection of sequences, RNNs are capable of generating new sequences. Speech synthesis models typically generate continuous outputs. RNNs have been used to predict speech feature vectors at each time step, which can be directly used as input to a vocoder for output. Moreover, RNNs can produce an optimal set of phonemes given an input text, where the output audio is then voice-converted to a better-sounding speaker. RNNs used in the current attention-based systems outperform earlier DNN architectures. The LSTM and GRU adaptations have regular states that help in learning the sequences, and additional gates to control state information, have both been successful to mitigate issues of vanishing and exploding gradients in RNN training. Training RNN models for speech is difficult, and that successful training requires many considerations to be taken into account. Finally, there remain a number of challenges and considerations for deploying RNN models in speech systems.

4.2. Convolutional Neural Networks (CNNs)

The transformer is a model architecture that has shown ground-breaking performance in a number of natural language processing tasks. So, do not omit being this sample, try to report on it. In fact, the transformer cannot be omitted. However, not to be disqualified as plagiarism, this text suggests showing that attention-transformer models are the state-of-the-art awarded models recently, and the final pipeline is an attention-transformer model.

4.3. Transformer Models

Transformer models are introduced as a crucial and revolutionary architecture in the landscape of speech synthesis, thanks to great progress made in natural language processing. The guiding principles underlying transformers, including self-attention mechanisms, are comprehensively introduced. Various voice-based transformers and transformer-inspired algorithms in speech applications are discussed and well summarized. Transformer models have substantially outperformed many other models in speech synthesis tasks and have shown great potential in generating coherent and more contextually relevant speech (Li et al., 2018). Built upon the self-attention mechanism, such transformer models can more effectively learn contextual information between speech segments and can better generate speech highly relevant to input text. Encoder-decoder structured transformer models have also gradually become the dominant front-end design strategy for speech synthesizers to better learn acoustic information from generated utterances. Transformer models have also fueled a growing interest in utilizing pre-trained models in different forms to advance speech synthesis models progressively. (Guo, 2022)

SR-TTS: a rhyme-based end-to-end speech synthesis system is introduced. Rhyme is an important and distinctive characteristic in Chinese poetry, which refers to a pithy core-word sentence pattern. By decoding a sentence to a poem, a rhyme-constrained training strategy that uses the Hidden Markov Model was implemented to alleviate the issue of non-phonetic symbol sequences. Techniques like rhyme embedding and simplex rhyme auxiliary training were subsequently improved to benefit rhyme modeling in a transformer-based TTS model (Yao et al., 2022). Through an extensive exploration of rhyme score calculation, several search strategies were adopted to encourage the TTS model to generate more rhymed output. A carefully designed

dataset and rhyme evaluation toolkit were also presented to promote further research. While bringing many significant benefits, a transformer model's ability to capture a large number of dependencies also poses a great challenge in training time complexity. Meanwhile, a new open-source toolkit with precise enhancement in BERT adaptation and training speed can reduce memory requirements to smoothly train models. This article aims to explore relevant works on transformers from a dedicated emphasis on voice and speech in response to the rapid developments in this field and its growing importance and potential for the audio-domain.

5. Applications and Future Directions

Advances in the fields of machine learning and computational phonetics in the last few decades have facilitated enormous growth in the field of computer speech processing. The development of current applications and considerations of future impact of this technology have potential relevance to a wide range of disciplines and industries, and the search for text-to-speech (TTS) systems or automatic speech recognition (ASR) systems has generated a growing interest in the system and computer phonetics of human speech. Use of TTS in telecommunications and the entertainment industry is rapidly growing. Accessibility to the user is enhanced through the TTS use of mobile phone functionality, emailing for fixed text, such as emails and web pages, and accessing programme guides for people in any area or advertising (Cordella et al.2022). Similarly, the intelligent personal assistant industry has been enabled by devices such as Amazon Echo, Apple's Siri and Microsoft's Cortana utilizing technology. Intelligent personal assistants offer a number of services to the consumer, such as checking the weather, reading out calendars, making phone calls etc., or the ability to access other IoT devices of the consumer (e.g. the ability to turn on the lights or heating). This is just one of the main applications of ASR in the field of distributed interfaces and swiftly growing. In the field of healthcare remarkable use of ASR technology are the international audio-directed diagnostic systems, saving valuable time for the identification of a patient's condition, and the auditory guiding system, which facilitate surgical procedures that are cost-effective, when communications between surgeons are obstructed or troubled by sound constraints or other barriers (Denby et al., 2022). Speech technology is now also flourishing in a wide range of other applications, such as dialogue

interaction systems, language learning applications, linguistic analysis and indexing of spoken corpora and animations and avatars to improve speech work communication or any other communities with loud noise in the premises. However, there is evidence supporting the TTS systems that a significant part of the scientific, legislative and professional accent is paid. This, at a basic level, mirrors the fact that there is an increasing acceptance that the success of speaking dialogue systems (SDS) depends on the way the ASR and TTS systems are made, including that it is intuitive and so human-like in its speech. From the technical point of view, this means that a lot of systematic collaboration and system engineering is required between these components, and more broadly, a significant part of the future impact of ASR and TTS technologies will relate to the way in which they are made. There have been a number of strategies over the past decade or so where machine learning methodology has been employed to enhance the naturalness and expressiveness of synthetic speech. Importantly, a large part of this work has been driven by the construction of novel corpora having expressive variants of the same phonetic unit either from the same speaker or different speakers. Since the expressive variants are labeled, this makes the task of the selection of the expression levels far easier than under optimal examination situations. One of these corpora has been used together for a range of analysis techniques to investigate expressive speech. While the overall focus of these studies has been on our understanding of the phonetic, prosodic and physical characteristics that distinguish different expressions in speech, much of the methodology, particularly that applied to prosody, has been used to explore the robustness of different parameters as cues to the expression level in the corpus. There has been a proliferation of projects associated with expressive speech synthesis, understanding emotions in speech, and even applications to dancing and contemporary bodies. Obviously, there is a simultaneous concern for the future that technology should not be over-regulated and constrained so that growth, especially in small scale research environments, is stifled. However, it is important to be responsible in the way these new systems are used and to be aware of the likely societal and ethical implications. This will affect how ASR and TTS technologies are already being deployed and will be exploited far into the future. The ethical issues divided these include the potential impact of societal fear and how the society may exploit the technology. They also consider how this kind of fear might be inbuilt into the design of these applications in a socially

responsible way. It is suggested that the impact of speech technology may not be as revolutionary as some expect as the potential ethical concerns of society will inhibit deployment. Other societal needs and fears are likely to drive different applications and will vary from technology to technology, and it is also important to consider methodological privacy and what kind of information methods might be investigated or used suspectably, this will include a discussion of concerns with large scale monitoring, ergonomic experimentation. However, this text also aims to consider future dialogue connection systems (SDS) more widely and future challenges and new investigations are divided into three main sections. In section 2, some ongoing research studies on expressing synthetic speech is accounted for and in the case of TTS technologies. This is followed by a consideration of potential future challenges and new studies are considered how different speech intelligent environments are likely to grow and begin to merge with one or the of the inclusive findings of inquiry described.

References:

Al-Fraihat, D., Sharrab, Y., Alzyoud, F., Qahmash, A., Tarawneh, M., & Maaita, A. (2022). Speech recognition utilizing deep learning: A systematic review of the latest developments. Human-centric Computing and Information Sciences, 14. researchgate.net

Bhangale, K. B., & Mohanaprasad, K. (2021). A review on speech processing using machine learning paradigm. International Journal of Speech Technology, 24(2), 367-388. pccoer.com

Brahmi, Z., Mahyoob, M., Al-Sarem, M., Algaraady, J., Bousselmi, K., & Alblwi, A. (2022). Exploring the Role of Machine Learning in Diagnosing and Treating Speech Disorders: A Systematic Literature Review. Psychology Research and Behavior Management, 2205-2232. tandfonline.com

Cordella, C., Marte, M. J., Liu, H., & Kiran, S. (2022). An introduction to machine learning for speech-language pathologists: concepts, terminology, and emerging applications. Perspectives of the ASHA Special Interest Groups, 1-19. asha.org

Denby, B., Gábor Csapó, T., & Wand, M. (2022). Future Speech Interfaces with Sensors and Machine Intelligence. ncbi.nlm.nih.gov

Guo, J. (2022). Innovative Application of Sensor Combined with Speech Recognition Technology in College English Education in the Context of Artificial Intelligence. Journal of Sensors. wiley.com

Iskarous, K. & Pouplier, M. (2022). Advancements of phonetics in the 21st century: A critical appraisal of time and space in Articulatory Phonology. Journal of Phonetics. [HTML]

Jain, N. & Rastogi, S. (2019). SPEECH RECOGNITION SYSTEMS – A COMPREHENSIVE STUDY OF CONCEPTS AND MECHANISM. [PDF]

Karagthala, J. J. & Shah, V. (2022). Analyzing the recent advancements for Speech Recognition using Machine Learning: A Systematic Literature Analysis. Journal of Electrical Systems. [HTML]

Kumar, Y., Koul, A., & Singh, C. (2022). A deep learning approaches in text-to-speech system: a systematic review and recent research perspective. Multimedia Tools and Applications. [HTML]

Latif, S., Rana, R., Khalifa, S., Jurdak, R., Qadir, J., & W. Schuller, B. (2020). Deep Representation Learning in Speech Processing: Challenges, Recent Advances, and Future Trends. [PDF]

Li, N., Liu, S., Liu, Y., Zhao, S., Liu, M., & Zhou, M. (2018). Neural Speech Synthesis with Transformer Network. [PDF]

Mehrish, A., Majumder, N., Bharadwaj, R., Mihalcea, R., & Poria, S. (2022). A review of deep learning techniques for speech processing. Information Fusion, 99, 101869. sciencedirect.com

Nagesh Shewalkar, A. (2018). Comparison of RNN, LSTM and GRU on Speech Recognition Data. [PDF]

Reflinda, R., Roza, V., & Firdaus, F. (2022). Exploring Emerging Trends in Phonetics: The Influence of Orthographic Forms and Technological Integration in Language Learning. REiLA: Journal of Research and Innovation in Language, 6(2), 193-206. unilak.ac.id

Tan, X., Qin, T., Soong, F., & Liu, T. Y. (2021). A Survey on Neural Speech Synthesis. [PDF]

Tits, N., El Haddad, K., & Dutoit, T. (2019). The Theory behind Controllable Expressive Speech Synthesis: a Cross-disciplinary Approach. [PDF]

Yao, Y., Liang, T., Feng, R., Shi, K., Yu, J., Wang, W., & Li, J. (2022). SR-TTS: a rhyme-based end-to-end speech synthesis system. ncbi.nlm.nih.gov