International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 40, No. 5, October, 2023, pp. 1997-2008
Journal homepage: http://iieta.org/journals/ts

A Comprehensive Examination of Phoneme Recognition in Automatic Speech Recognition Systems

Shobha Bhatt[1], Shweta Bansal[2], Ankit Kumar[3], Saroj Kumar Pandey[3], Manoj Kumar Ojha[2], Kamred Udham Singh[4], Sanjay Chakraborty[5], Teekam Singh[6], Chetan Swarup[7*]

[1] Department of Computer Science and Engineering, Netaji Subhash University of Technology, Delhi 110078, India
[2] Department of Computer Science and Engineering, K.R. Mangalam University, Gurugram 122103, India
[3] Department of Computer Engineering & Applications, GLA University, Mathura 281406, India
[4] School of Computing, Graphic Era Hill University, Dehradun 248002, India
[5] Department of Computer Science and Engineering, Techno International New Town, Kolkata 700156, India
[6] Department of Computer Science and Engineering, Graphic Era Deemed to be University Dehradun 248002, India
[7] Department of Basic Science, College of Science and Theoretical Studies, Saudi Electronic University, Riyadh-Male Campus, Riyadh 13316, Saudi Arabia

Corresponding Author Email: c.swarup@seu.edu.sa

https://doi.org/10.18280/ts.400518

Received: 9 January 2023
Revised: 5 March 2023
Accepted: 15 March 2023
Available online: 30 October 2023

Keywords:
review phoneme, speech, consonants, vowels recognition

ABSTRACT

This review offers an exhaustive examination of phoneme recognition, an essential sub-word acoustic unit in speech processing. Phoneme-based systems find widespread utility in diverse applications including speech recognition, speaker identification, and language recognition. The efficacy of these systems hinges upon the precise recognition of phonemes, thereby underscoring the criticality of enhancing our understanding of phoneme recognition to optimize system performance. Previous reviews have primarily focused on specific issues within the realm of phoneme recognition, with comprehensive studies on the subject being notably sparse in existing literature. Consequently, there is an urgent need for an extensive investigation into phoneme recognition to bolster recognition accuracy. This comprehensive review seeks to bridge this knowledge gap by examining pivotal aspects such as vowel recognition, consonant recognition, acoustic-phonetic cues, contextual effects, feature extraction methods, classification techniques, phoneme recognition enhancement strategies, and performance metrics. The review elucidates various technologies and trends in phoneme recognition, thereby providing valuable insights that can mitigate errors in phoneme-based systems through the application of appropriate techniques delineated in the study. The findings of this study hold substantial potential benefits for a wide spectrum of speech research communities, encompassing students, educators, specialists, developers, and scholars. The review encompasses both fundamental and advanced concepts pertinent to phoneme recognition, thereby offering a comprehensive resource for individuals engaged in this field.

1. INTRODUCTION

Automatic speech recognition (ASR) systems often utilize sub-word models to circumvent the issue of insufficient data during training that plagues many word-based models. A robust sub-word model should possess an ample number of instances to successfully model word utterances, while also being resilient to environmental variability [1]. Among these, phoneme-based systems have gained prominence in speech recognition due to the limited number of phonemes in any given language, facilitating the application of various operations and rules to form words through the combination of phonemes [2, 3].

However, phoneme-based ASR systems encounter numerous challenges, including phoneme confusion, variations in speaking rates and styles, and inherent variability in the speech signal. Context-dependent triphones have been employed to address these contextual effects [4-7].

Phonemes can be broadly classified into two categories: vowels and consonants. Efforts have been made to mitigate the contextual effects on vowels by considering factors such as vowel identity, identities of adjacent segments, syllable placement within a word, vowel placement within a syllable, stress status of a syllable, effects of phrase boundaries, and word accent status [8]. Larger acoustic units were also utilized to lessen the influence of contextual factors [9]. Consonants face recognition difficulties due to the manner in which they are produced, and semivowels pose additional recognition challenges due to their acoustical similarity to vowels [10]. Furthermore, consonants can be sub-divided based on place and manner of articulation. Recognizing nasalized sounds is particularly challenging due to their anti-resonance properties. Phoneme recognition is also problematic when the crucial cue is duration, as is the case with vowels and diphthongs. To address duration, researchers have employed time-delay neural networks (TDNNs) for these phoneme classifications [11].

Phoneme recognition systems are expected to cover a broad range of contextual variations. Data scarcity becomes an issue during phoneme recognition when the training data for some phonemes are inadequate. Knowledge-based systems have been explored to mitigate the impact of data scarcity and mismatches between training and testing conditions [12]. Variability issues have been addressed using linguistic phonetic knowledge [13].

Previous reviews on the subject of phoneme recognition have largely focused on limited or specific issues [14-17]. Comprehensive reviews on phoneme recognition are notably lacking in the literature. This research review aims to bridge this gap by offering a comprehensive examination of phoneme recognition, including fundamental concepts and important issues such as phoneme structure, vowel recognition, consonant recognition, acoustic-phonetic cues, contextual effects, feature extraction methods, classification methods, phoneme recognition improvement techniques, and performance measurement metrics.

This research contributes by investigating the following aspects of phoneme recognition:

1. Exploration of vowel and consonant recognition.
2. Examination and comparison of different acoustic-phonetic measures used in phoneme recognition.
3. Exploration of contextual effects and improvement techniques in phoneme recognition.
4. Investigation of classification methods, acoustic-phonetic approaches, and feature extraction methods.
5. Exploration of performance metrics.
6. Presentation of a comparative study of phoneme-based systems.

The remainder of the paper is structured as follows: Section 2 reviews related work. Sections 3 and 4 delve into phoneme and phoneme recognition respectively. Section 5 investigates acoustic-phonetic measurements in phoneme recognition. Section 6 describes contextual effects in phoneme recognition. Section 7 illustrates the methods for improving phoneme recognition. Sections 8 and 9 outline performance matrices, results, and analysis. Finally, Section 10 concludes the research findings and provides directions for future work.

2. RELATED WORK

Phoneme recognition is an essential step towards developing speech-based applications due to the natural way of interacting with machines through speech. Researchers have worked actively towards phoneme recognition for developing different speech based applications. A review study on Hindi phoneme recognition was presented [18]. Research findings reveal that phoneme recognition was implemented using different classification methods: hidden Markova models (HMMs), artificial neural networks (ANNs), time delay neural networks, and deep learning-based systems. To improve recognition, different feature extraction methods were also applied. Most of the work was presented for vowel recognition. The substitution error was mostly found in the studies.

A study on Hindi phoneme recognition was conducted. Further different classification methods and feature extraction methods for Hindi speech were also explored. It was concluded that most of the research work was oriented toward vowel recognition, and errors were reported due to the substitution of the phonemes [18].

A review was conducted for recognizing phonemes using three classifiers Hidden Markov Model (HMM), artificial neural network (ANN), and vector quantization with comparative analysis [19]. The review indicated that phoneme recognitions were performed on continuous speech, isolated words, and rhythmic words. Different databases such as TIMIT, Hindi rhythmic words, and individual phoneme sets were applied. The classification algorithm using statistical methods, hybrid approaches, RNN, K-means, neural network, and vector quantization in the experiments were applied. Remarks were provided that global optimization of ANNs improved results, and sequence learning due to the feedback mechanism also performs significantly. Further, segmental HMM used phoneme transition data and phoneme length information.

A survey was conducted on TIMIT phonemes for deep learning-based approaches. The study included different research work using a feed-forward network with rectified linear units (ReLU), Time Delay neural network (TDNN), and long short-term memory (LSTM) neural networks. The best-reported phone error rate (PER) is 15.73%, with LSTM with five hidden layers and 512 LSTM units, while TDNN reported PER of 16.91%. It was observed that the best results were obtained with LSTM [15].

A class of deep artificial neural networks containing a minimum of three layers was used. The Bangla phonetic feature table was constructed. The improvements in Bengali speech recognition were also discussed. A comparative analysis of these two methods was also presented [20]. An elaborate and comprehensive review of the acoustic-phonetic assessment of speech and its use in ASR was presented [14]. Firstly, essential cues for recognition were discussed. Next, the implicit acoustic-phonetic and explicit phonetic methods to ASR are highlighted. The ANNs, HMMs, and dynamic time warping (DTW) are examples of implicit approaches to speech recognition. The two essential phases are training and testing for implicit strategies using statistical methods for ASR systems. The explicit approaches to recognition are based on specific knowledge. Some examples are landmark-based speech recognition and event-based speech recognition, such as combinations of vowels and consonants such as CV/CVC/VC/CCVC. Finally, different speech recognition frameworks were presented and compared.

A study on landmarks-based based stop consonants recognition was conducted. It was remarked that stop consonants are generated due to the constriction of the vocal cord; therefore, they pose difficulty in detection due to low energy and high variability. The study revealed that researchers used landmarks that are denoted by abrupt and essential changes in articulation. Further, it was concluded that VOT is a vital landmark for separating voiced and unvoiced stops. These two landmarks can also be applied to the problem of dysarthria, which is caused by the disorder of the speech production system [17].

The genetic algorithm was applied to select optimized feature vectors for distinctive phonetic features and phoneme recognition in Arabic to reduce the recognition algorithm's computational overhead and improve recognition accuracy. It was concluded that genetic algorithm-based features reduced the dimension of the input vector by 50% and obtained a recognition accuracy of 90% [21].

Persian Speech recognition was implemented using deep learning methods [22]. Faddat Persian speech dataset was applied for developing the speech recognition system. The features were extracted using Deep Belief Network (DBN). The acoustic model was generated using Deep Bidirectional

Long Short-Term Memory (DBLSTM) with Connectionist Temporal Classification (CTC) at the output layer. It was concluded that a bidirectional network increases accuracy compared to a unidirectional network, and the DBLSTM-based system performed better than HMM and Kaldi_DNN.

Studies also indicate that machine learning-based algorithms are preferred to the acoustic phonetic-based algorithm for the language having resources for developing speech recognition systems [7]. The automatic phoneme recognition system was developed for Fongbe under-resourced tonal language spoken in Africa. The read speech corpus contained 3117 utterances. For acoustic analysis, formant frequencies were applied for vowels, while pitch and intensity were used for consonants. For classification deep belief network with fuzzy logic was experimented with for phoneme recognition. The best results were obtained with a 24% phoneme error rate with 512 hidden layer units [23].

Acoustic to phonetic conversion experimented for Arabic speech. The distinctive phonetic features that particularly indicate a phoneme's distinctive quality were applied. The spectrogram was used for distinctive phonetic features (DPF). Deep learning methods such as deep belief networks (DBN) and convolution recurrent neural networks (CRNN)were applied for classification. The results indicate that CRNN is providing better results than DBN for DPF [24]. Bengali vowel and diphthong recognition was proposed by using amplitude interpolation. The Mel Frequency Cepstral Coefficients were applied. Different machine learning-based classifiers were applied [25].

To take advantage of ANN-based processing the Arabic phoneme recognition was implemented. For feature extraction multi-wavelet transform was applied. For classification Learning Vector Quantization (LVQ) was applied. It was stated that the developed system obtained accuracy of 98%.

To address the issue of phoneme recognition in spoken term detection for low-resource Indian language like Marathi, Gujrati. Malayalam, and Kannada language. The MFCCs were extracted. The multilingual broad phoneme classifier was used to conduct language independent spoken term detection. In the first phase broad phoneme classification was performed and then in the next phase template matching was performed. DNN based broad phoneme classifier was used. The broad classification based on categories such as vowel category, nasal sound category, fricatives, silent category, approximants, affricates, voiced plosives category and unvoiced plosives was conducted, it was concluded that the use of broad phoneme classifier is capable of independent spoken term detection for low resource languages.

Research studies show many challenges for phoneme recognition. Figure 1 shows a summarized view of different problems faced during the development of phoneme recognition systems. The literature review revealed that phoneme recognition is influenced by duration of the phoneme, speaking rate, style, accent, contextual effects, age, gender, health condition, training and testing environments, confusion of phonemes within the same categories and lack of state-of-the-art resources. There is a need to investigate phoneme recognition to improve speech recognition as phonemes are basic in automatic speech recognition systems and accurate recognition of phonemes leads to improved speech recognition.

The work presented in this study is different from others. All the above research reviews address specific issues in phoneme recognition. The presented work covers broad aspects of phoneme recognition, which are essential for reducing errors and improving phoneme recognition. The presented study also addressed basic and advanced issues in phoneme recognition. The basic concepts in phoneme recognition were discussed to enable the reader to understand the topic. Different challenges were discussed and reviewed to address phoneme recognition.

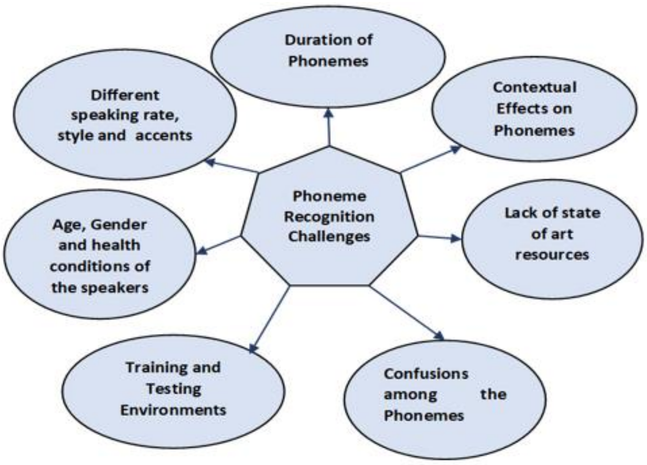

Duration of Phonemes
Different speaking rate, style and accents
Contextual Effects on Phonemes
Phoneme Recognition Challenges
Age, Gender and health conditions of the speakers
Lack of state of art resources
Training and Testing Environments
Confusions among the Phonemes

Figure 1. Various challenges in phoneme recognition

3. SUBWORD MODELLING UNITS

The subword models are used in speech recognition in place of word-based models to overcome the requirement of large instances of words during the training phase. The words are made from these sub words for the realization of a speech recognition system. The choice of sub word modelling unit and the methods to generate words from these sub word units is the focus of sub word modelling in speech recognition. Figure 2 shows the taxonomy of different sub word modelling units used in speech recognition [26].

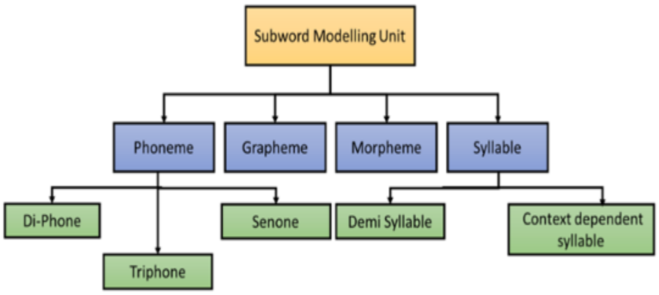


Figure 2. Different subword modeling units in automatic speech recognition systems

The subword modelling units can be categorized as phoneme, grapheme, morpheme, and syllable. The phoneme is the smallest unit of sound speech, while grapheme is the smallest writing unit in a language [27]. The morpheme is the smallest unit of meaning [28]. The syllable is the unit of pronunciation. A phoneme can be further divided into di-phone and senone. The di-phone is the combination of two adjacent phonemes in an utterance. The senone are defined as tied states in context-dependent phonemes [29]. The context-dependent phonemes such as triphones are generating by using the left and right contexts of the phoneme. The demi-syllable is made of an initial syllable consonant cluster with the first half of the vowel or syllable-final consonant cluster plus the second half of the vowel [30]. The context-dependent syllable is realized by considering the left and right context of the syllable [31].

3.1 Phoneme

To understand phoneme recognition basic process of creating phonemes and related issues are essential to explore and study. The human speech production system generates speech sounds. The two most important systems vocal tract and the nasal cavity, describe speech production phenomena in speech processing. The articulators move to produce different sounds. Different formant frequencies and antiresonance principles narrate acoustic properties for different speech sounds. The vibration generates the voiced speech in the vocal tract. The turbulent airflow causes unvoiced speech due to constriction in the vocal tract. Phonetics deals with the structure of sounds, and linguistics deal with creating rules for converting sounds into information. The phoneme is the fundamental unit of speech with defined numbers in every language.

The consonant and vowels are the main parts of phonemes. There are 42 phonemes containing vowels, semivowels, diphthongs, and consonants in American English. The consonants include nasals, stops, fricatives, and affricates. The vowels are produced by vocal fold vibration. Other categories similar to vowels are diphthongs and semivowels. A diphthong is a transition from one vowel to another. The semivowel is further classified into liquid and glides. Further consonants are produced by constriction of the vocal tract. The fricatives are generated by excitation of the vocal tract with a constant airflow that becomes turbulent due to constriction. The fricatives may have voicing components, therefore called voicing fricatives. Affricatives are produced by the transition of a stop to fricatives. The stop consonants are produced by creating pressure behind the full constriction of the vocal tract and abruptly releasing the pressure. The nasals are voiced consonants produced by an exciting nasal cavity. Figure 3 shows the phoneme structure used in the well-known TIMIT speech corpus [32].

At the first level of the hierarchy, the phoneme structure is categorized into obstruent, silences and sonorants. The obstruents are generated with turbulent noise, such as fricative, plosives and affricatives, as shown in the subcategory of the obstruent in Figure 3. The sonorants are generated without turbulent noise and divided further into vowels, nasals and liquids. The fricative and plosives are further divided into voiced and unvoiced sounds, and vowels can be further divided into back, front and centre vowels. The TIMIT speech corpus is a dataset of recorded speech that is widely used in speech processing and recognition research. It contains phonetically transcribed recordings of 630 speakers, representing eight major dialects of American English. The phonetic structure of the TIMIT corpus is based on the International Phonetic Alphabet (IPA), which is a standardized system for representing the sounds of human language. The corpus includes phonetic transcriptions at the level of individual phonemes, as well as at the level of phonetic features such as voicing, place of articulation, and manner of articulation. The corpus is organized into training, development, and test sets, which are used to train and evaluate speech recognition systems. The phonetic structure of the corpus is carefully designed to ensure that the training, development, and test sets are representative of the range of phonetic variability in natural speech.

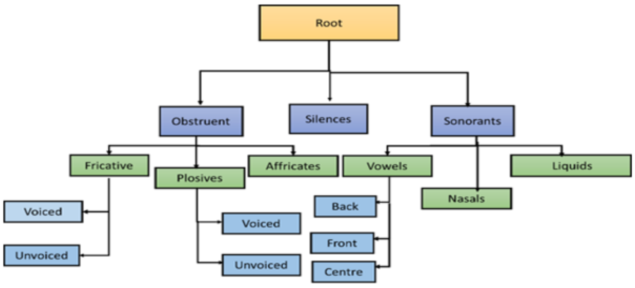


Figure 3. TIMIT speech corpus phonetic structure

Further language issues also play an important role in recognition. Figure 4 shows Pierce's model for natural language construct [33-36].

The model contains four constructs. These constructs are symbolics, grammatical, semantic, and pragmatics. All these constructs also have prominence in speech recognition. The symbols may be any speech unit such as word, phoneme, and syllable.

Pragmatics is concerned with studying sentence interpretation according to the contextual situation. Grammar is the study of the rules in a language. Semantics is related to learning and analysing the meaning of the text. The symbols may be any speech unit, such as word, phoneme, and syllable.

The idea that language is a system of signs that are used to represent things, ideas, and concepts is central to Peirce's model for natural language formation. As Peirce sees it, a sign has three parts: the signer, the signified, and the interpretant.

A phrase, gesture, or picture are all examples of signifiers. What a sign actually depicts, or the signified, is a meaning or an idea. The effect of the sign on the interpreter, known as the interpretant, might range from a straightforward comprehension of the sign's meaning to a tangled web of inferences and inferences.

In Peirce's theory, context plays a crucial part in establishing the meaning of signals. In particular, he contended that signals are not unchanging objects but rather take part in a continuous cycle of interpretation and mutual influence among the signifier, the signified, and the interpretant. So, a sign's meaning can shift depending on the circumstances in which it is employed and the perspective from which it is viewed.

Semiotics, the study of signs and symbols, owes much to Peirce's framework. Linguistics, communication studies, and philosophy are just some of the disciplines that have benefited from its application to questions of meaning and the connection between language and cognition and the world as we know it.

Further, grammar connects the symbols to form a message unit. Speech recognition scientists use pronunciation dictionaries to create lexicons to generate words from the specified symbols. However, grammar may create any word, so for creating meaningful words, semantics is used. Pragmatics deals with contextual effects in natural language understanding systems. It was stated that this is the hardest part to incorporate into the speech understanding system.

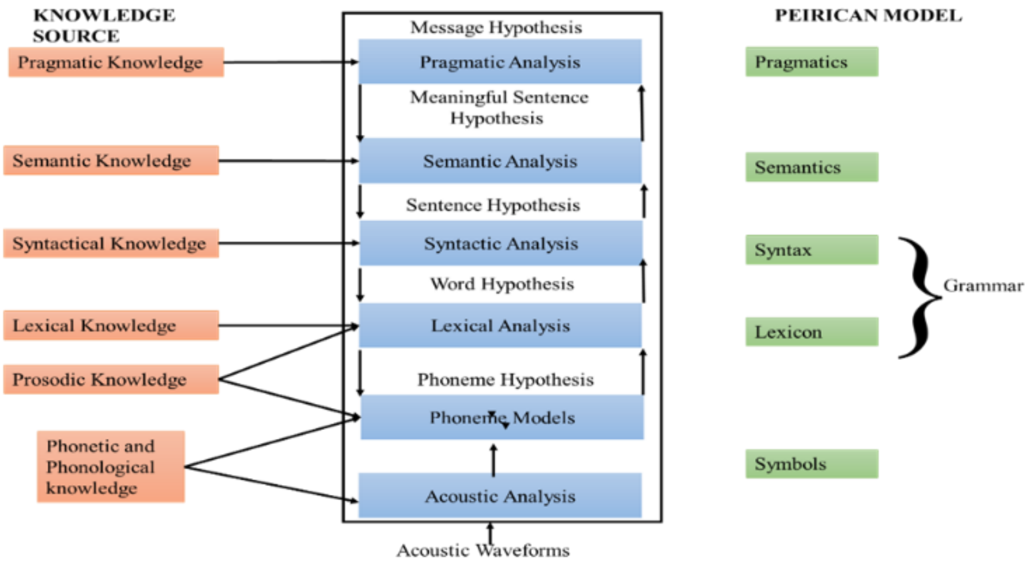


Figure 4. Peirce's model for natural language construct [33-36]

4. PHONEME RECOGNITION

Every speech recognition process starts with creating a speech text and developing a speech corpus. After that, relevant information is extracted in terms of features from the speech signal. Further, the phoneme-based acoustic models (AM) are created from the features, and the trained model is produced after using re-estimation. A highly probable word series is identified from the trained AM and language models (LM) with an ASR engine.

4.1 Vowel recognition

Vowel recognition using the formant was presented [37]. The speech corpus was recorded with 80 Malaysian speakers. The new feature extraction method using the bandwidth approach was applied. The vowels can be defined with stable frequency phenomena, and it sets a unique basis for recognition of the vowels due to the various average vocal tract length present in males, females, and children. The mean magnitudes were also calculated from the bandwidth frequency ranges. The results were also compared with 13 Mel Frequency Cepstral Coefficients (MFCC). The backpropagation neural network was used for classification.

The contextual effects on vowel duration were presented in [8]. The study was performed on two speech corpora consisting of 18000 and 6000 vowel segments from the spoken utterances by American English speakers. The research was conducted to investigate the new facts already available in the literature. Different factors such as stress and accent, statistical methods, piecewise multiplicative corrections, the sum of product models, and vowel identity were considered. It was concluded that intrinsic vowel duration, pitch accent, position, stress, post, and prevocalic consonants accounted for 86% of the variance. It was also discovered that pitch accent amplified the effect of syllabic stresses in accented words, while deaccented syllables were less affected. The postvocalic consonant cluster effects were related to voicing and the manner of production. Syllable boundaries also have a significant role in contextual effects. Investigations revealed that the length of the vowels becomes shorter in closed syllables.

4.2 Consonant recognition

Consonants constitute a significant part of any phoneme-based system. Efforts have been made to classify nasals and semivowels automatically [38]. TIMIT speech corpus was selected for the study. Different measures based on onset/offset were used for the nasal sounds. A support vector machine (SVM) classifier was used to combine acoustic properties. The results reported for different sonorant consonant categories are 88.6% for prevocalic, 94.9 for postvocalic, and 85.0% for intervocalic. The overall recognition score achieved by nasals was 92.4%, while semivowels reached 88.1%. The recognition of semivowel was presented by using linguistic features for American English. The features used were sonorant, syllabic consonantal, high, back, front, and retroflex. The speech corpora included polysyllabic words and sentences with various dialects. The recognition results were reported for semivowels [10].

The stop consonants in consonant-vowel (CV) words were recognized using acoustic-phonetic features [12]. The study used two speech corpora, TIMIT and NTIMIT (telephonic). Spectral processing based on FFT was applied. The acoustic cues such as the relative centre of gravity, burst amplitude, voice onset time, and log of ratios of first formant prominence were used as acoustic-phonetic features. The study aimed to distinguish labial /p/, alveolar/t/ and velar/k/. The best results reported are 85.7% for overall recognition, while the labial sound got 89.3% recognition, the alveolar sound got 84 % recognition, and the velar sound got an 84.8% recognition score.

5. ACOUSTIC PHONETIC MEASUREMENTS IN PHONEME RECOGNITION

Different acoustic-phonetic measurements were used for phoneme recognition, such as voiced and unvoiced parameters, articulatory features, formant-based measurements, burst and voice onset time, vowel offset and onset points, nasalization, and fricative. Phonetic reduction happens when a phoneme moves away from its natural form. It was concluded that vowel reduction is connected to the duration and quality of vowels

[39, 40]. The research findings also indicate that formant frequency is the most frequent parameter for forensic speaker practices [41]. Many parameters such as formant values, duration, mean frequency and energy differences were also studied. Acoustic phonetic variability was addressed for Polish vowels [13]. Three levels of abstraction, such as intrinsic allophonic, extrinsic allophony, and phonemic variation, were defined to describe the speech sound. The research work aimed to address intrinsic allophony issues in Polish vowels.

Table 1 shows the comparative analysis of the parameters used for usefulness in recognition, the method used for calculating these measurements, challenges, and improvements to these methods.

Table 1. Comparative analysis of different exiting method with different parameters [14, 42-47]

Acoustic-Phonetic Cues	Usefulness in Recognition	Methods Used	Challenges	Solution to Challenges
Voiced and unvoiced	Information regarding voicing is essential for phoneme recognition, as almost all stops are voiced and unvoiced. The confusion analysis is reduced due to voiced and unvoiced information among the same place of articulation.	The parameters used for voiced activity detection are zero-crossing rate, predictive coefficients, autocorrelation coefficients at first lag, and peak strength harmonic measures. Further, multiple features were combined with GMM, HMM, and ANN-based methods were also used.	Easy when voicing energy is high. Weak glottal activity weak parameters such as nasals are challenging to detect.	LP residual, dominant resonant frequencies, Perceptual linear predictive coefficients, and Wavelet-based features are used.
Formants based measurement	Formants are used in detecting vowels and are also useful for recognizing vowel-like regions, recognizing consonants-vowel, robust against channel distortion and noise and well suited to tackle mismatch between testing and training environments. Formant transition shows the place of constriction.	The formants are estimated by calculating and then tracking the formants. Different methods are centred on the Linear prediction and cepstrum assessment.	Peak-picking techniques in formant estimation are susceptible to joined formants and spurious peaks. Accurate calculation of formants is also an important task.	A set of digital formant resonators in parallel are used.
Burst and voice onset time	Voice onset time (VOT) includes helpful information about consonants related to articulation; the smallest VOT is observed for bilabial and grows steadily in the direction of velar sounds.	The maximum value of the normalized cross-correlation calculated between consecutive inter epoch periods, thresh hold logic for detection of stop burst, fuzzy classifier, and rule-based classifier.	The complications are caused due to the effect of place of articulation on VOT values.	Language-specific phonological rules assigned and multitasking learning are used.
Nasalization	Useful phoneme recognition and lowers misunderstanding between nasalized and non-nasalized vowels.	It is characterized by the first two resonance frequencies, the existence of additional peaks: one between the first two formants and one at lower frequencies for nasalized vowels. Low to high order residual energy ratio and dominant resonance for nasal murmur recognition.	Automatic detection of antiresonance is difficult.	Detecting nasality relied mainly on robust formant extraction and feature extraction methods such as PLPs.
Fricative	Acoustic phonetic information is useful in phoneme recognition. Recognition of unvoiced fricatives from voiced fricatives increases recognition accuracy.	The parameters for calculation of fricatives such as the zero-crossing rate for voiceless frication, LP spectrum, articulation place, and voicing event.	Detection of voiced fricatives is not easier and causes confusion between voiced fricatives and low energy sonorants such as semivowels and nasals.	Spectral features calculated from a short time frame using SVM show the improvements.
Articulatory features	The degree of constriction determines vowel highness, aspiration, and frication. Articulatory features are grouped into place, manner, voicing roundness, frontness, and height.	The methods, such as acoustic articulatory transformation using inverse mapping, direct physical measurement, and classification score for pseudo-articulatory features, are used to find articulatory features.	The inverse mapping is complicated in continuous speech. Direct physical measurement requires costly setup such as X-Ray filming, electromagnetic articulography, and electropalatography. The databases based on physical measurements for Indian languages are not available.	Most of the work is based on a spectral feature to derive articulatory features.

Vowel onset and offset points	The voice onset point (VOP) and voice endpoint (VEP) are essential in consonant recognition.	VOP measurements are based on zero-crossing rate, spectral peak, and ANN-based methods.	Difficulty in the detection of semivowel and voiced aspirated sounds.	To improve the accuracy amplitude envelope of the vowel emphasized amplitude, the modulated-frequency modula-ted signal was used for VOP and VEP.

6. CONTEXTUAL EFFECTS IN PHONEME RECOGNITION

The researchers extensively studied Hindi phoneme confusion analysis [48]. The contextual effects on vowel duration were obtained in the study [8]. It was concluded that syllable boundaries also significantly affect contextual effects. Context-dependent triphones were used to address the contextual effects [49]. The bidirectional Long Short Term Memory (BiLSTM) network was used for phoneme recognition [50]. It was concluded that the bidirectional system presents the input into two forward and backward networks, which are well fitted for addressing contextual effects. The phoneme recognition was addressed by using the HMM-ANN paradigm with contextual information [51]. The contextual effects were studied at the feature level and output level of the MLP.

The hierarchical phoneme recognition was implemented to address phoneme confusion within the same broad categories. Different levels of hierarchy were defined. First-level vowels and consonants were defined. Different categories such as semivowel, fricative, affricates, nasals, stop consonant category I and stop consonants category II were used at the second level. A further division was made within these categories. MFCCs were extracted for feature, and a support vector machine was used for classification. The experiment was conducted using the TIMIT database. The significant improvements in TIMIT speech corpus phoneme recognition were observed by applying hierarchical phoneme recognition compared to traditional speech recognition [52].

7. PHONEME RECOGNITION IMPROVEMENT TECHNIQUES

The researchers have suggested different techniques to improve phoneme recognition. Contextual effects play an important role in phoneme recognition. Researchers used triphone-based context-dependent phonemes to reduce the contextual effect [4, 53]. Many statistical parameters are related to phonemes, such as length, duration, and frequency. The paper addresses finding helpful statistical parameters in phoneme recognition. To improve Arabic phoneme recognition, helpful statistics based on HMM states for different phonemes were presented. It was concluded that various measures such as phoneme duration, frequency, and the probability of different phonemes occurring are helpful statistics to improve phoneme recognition.

Further, it was indicated that appropriate states could be designed as per the phoneme [54]. The small-duration phonemes suffer from deletion errors, so a proper feature extraction method is needed for small-duration phonemes. The researchers also applied wavelet-based features to improve phoneme recognition [55] to address small-duration phonemes. Researchers used syllable-based speech recognition to reduce the contextual effects [56]. Articulatory features were also applied to improve phoneme recognition [47].

Every classification method has its own advantages. Therefore, the combination of different classifiers can also lead to improve phoneme recognition. The phoneme recognition was improved using the HMM-ANN paradigm with contextual information [45]. The contextual effects were examined at the feature level and output level of the multilayer perceptron (MLP). The sub-phonemic groups were addressed at the feature level. The hierarchical estimate of phoneme posterior probabilities was intended to tackle contextual information at the output of MLP, and silence was excluded for the recognition accuracy measurement.

The research was conducted to include contextual information by applying the BiLSTM network that uses past and present information to improve phoneme recognition. Two experiments were performed using a TIMIT speech corpus with unidirectional and bidirectional LSTM networks. The first experiment was conducted for frame-wise phoneme classification. The results indicated that bidirectional LSTM performed better than unidirectional LSTM and conventional recurrent neural networks (RNNs). The second experiment was conducted using a combination of bidirectional LSTM and HMM. The system outperformed both traditional HMM and unidirectional LSTM-HMM. It was also stated that RNNs are useful when the span of contextual effects is known. Bidirectional systems present the input into two forward and backward networks, better suited for modelling contextual effects [44].

The recent works based on deep neural networks (DNNs) systems show improved speech recognition results [57].

DNN-based systems have the training and testing phases. Though earlier works improved the training phase, this work focused on the test phase. The research work was experimented using a real duration probability distribution for each phoneme using a hidden semi-Markov model (HSMM) instead of geometric distribution of state duration in HMM. Each phoneme was represented by only one state by simply using phoneme duration in HSMM. The researchers also examined the performance of a post-processing method that connects the phoneme sequence obtained from the neural network. Bigram language models with MFCCs were used in the experiments. Experiments were conducted on the Persian speech corpus. The results show that the extended Viterbi algorithm on HSMM improves phoneme recognition accuracy by 2.68% and 0.56% over the conventional methods using GMM-HMM and Viterbi on HMM, respectively.

8. PHONEME RECOGNITION PERFORMANCE EVALUATION

The studies show phoneme error rate (PER), phoneme accuracy, and phoneme correctness mostly used evaluation metrics. The PER is defined below [58-60].

$$\text{Phoneme accuracy} = (N\text{-}S\text{-}I\text{-}D)/N \quad (1)$$

$$\text{Phoneme error rate (PER)} = (S+I+D)/N \quad (2)$$

The following explains the variables used N, S, I, and D in Eqs. (1) & (2).

N: denote a total number of phonemes in the reference string.
S: indicates substitution.
I: indicates insertion.
D: indicates deletion.

9. RESULTS AND ANALYSIS

The HMM-based, ANN-based, GMM-based, dynamic time warping (DTW), and vector quantization based phoneme recognition were performed. The classification algorithm reported, such as HMM/ANN hybrid, takes advantage of the combined methods. The neural networks supervised, unsupervised learning algorithms, Kohonen map, RNN, and K-means for phoneme recognition were applied. Deep learning-based TIMIT phoneme recognition was presented. The study included different research work using a feed-forward network with rectified linear units (ReLU), Time Delay neural network (TDNN), and long short-term memory (LSTM) neural networks. It was observed that the best results were obtained with LSTM. Figure 5 shows the phoneme error rate (PER%) indicated by different methods on the TIMIT speech corpus [15, 61]. Table 2 presents a comparative analysis of phoneme recognition of various methods.

The Mel Frequency Cepsral coefficients (MFCCs), linear prediction cepstral coefficients (LPCs), Perceptual Linear Prediction Coefficients (PLP), wavelet-based methods for short duration phoneme, and cepstral mean normalization based features were utilized for phoneme recognition. The bandwidth-based feature extraction improved vowel recognition. Spectral processing based on FFT was applied. The acoustic cues such as the relative centre of gravity, burst amplitude, voice onset time, and log of ratios of first formant prominence were used as acoustic-phonetic features to classify stop consonants in TIMIT and NTIMIT. The research findings indicate that MFCCs are widely used in extracted features.

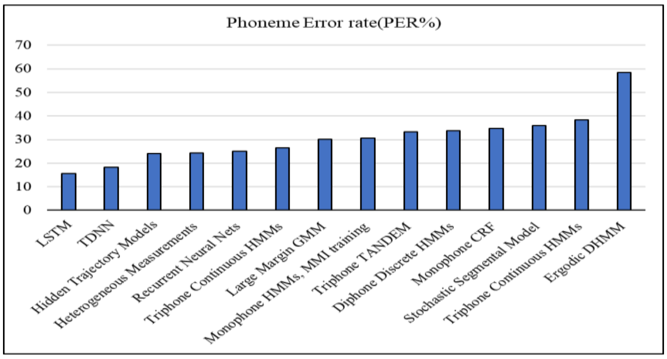


Figure 5. TIMIT phoneme recognition with PER%

Table 2. Comparative analysis for phoneme recognition

Reference	Features Used	Classification Method	Recognition	Accuracy
[38]	Onset/Offset	SVM	TIMIT consonants	Sonorant consonant categories 88.6% for prevocalic, 94.9 for postvocalic, 85.0% for intervocalic. The overall recognition score nasals: 92.4%, semivowels: 88.1%.
[12]	Spectral processing based on FFT was applied.	The acoustic cues such as the relative centre of gravity, burst amplitude, voice onset time, and log of ratios of first formant prominence were used as acoustic-phonetic features.	Voiceless stop consonants in consonant-vowel (CV) words. Two speech corpuses TIMIT and NTIMIT (telephonic) were used in the study.	85.7% for overall recognition, labial sound: 89.3% recognition score, alveolar sound: 84 % recognition score, velar sound got: 84.8% recognition score.
[37]	Bandwidth-based feature extraction.	Back Propagation Neural Networks	Malaysian Vowel recognition using the formant.	The bandwidth-based feature extraction provided the classification accuracy of 89.58%, which was 0.71% better than MFCCs.
[2]	MFCCs	HMM	Hindi vowel	Vowel:83.19%
[62]	MFCCs	GMM	Hindi vowel	Microphone recorded speech:91.4% Telephonic speech: 84.2%
[63]	Gammatone frequency cepstral coefficients and formant.	CDHMM	Hindi vowel	Speaker dependent: 99.15%, Speaker independent: 98.5%
[64]	MFCCs, along with vocal tract area function.	SVM, K nearest neighbours (KNN) and linear discriminant analysis (LDA) were combined with a voting classifier.	British English	Phoneme recognition accuracy: 83.95%
[57]	MFCCs	DNNs and HSMM	Persian speech corpus.	The extended Viterbi algorithm on HSMM achieves improvements in phoneme recognition accuracy of 2.68% and 0.56% over the conventional methods using GMM-HMM and Viterbi on HMM, respectively.

The study revealed that researchers used landmarks that are denoted by the abrupt and essential changes in articulation. Two prominent landmarks, VOT and burst release, were experimented with in the studies to detect stop consonants. Further, it was concluded that VOT is an important landmark for separating voiced and unvoiced stops. For vowel identification, these two landmarks can also be applied to the three articulatory features, high, backness, and roundness. Vowel recognition using the formant was presented. The vowels can be defined with stable frequency phenomena, and it sets a unique basis. The research was conducted to investigate the new facts which were already available in the literature. Different factors such as stress and accent, statistical methods, piecewise multiplicative corrections, the sum of product models, and vowel identity were considered. It was concluded that intrinsic vowel duration, pitch accent, position, stress, post, and prevocalic consonants accounted for 86% of the variance. It was also discovered that pitch accent amplified the effect of syllabic stresses in accented words, while deaccented syllables were less affected.

The categorization of consonants was made on the basis of place of articulation, manner of articulation, and voicing. The recognition of consonants is a challenging task in speech recognition due to their production methods. Researchers made efforts to recognize the consonants using different techniques in the literature. Efforts have been made to classify nasals and semivowels automatically. TIMIT speech corpus was selected for the study. Various measures based on onset/offset were used to capture the consonantal nature of the nasal sounds. A support vector machine (SVM) classifier was used to combine acoustic properties. The recognition of semivowel was presented by using linguistic features for American English. The used features were related to sonorant, syllabic consonantal, high, back, front, and retroflex. The research work aimed to identify voiceless stop consonants in consonant-vowel (CV) words.

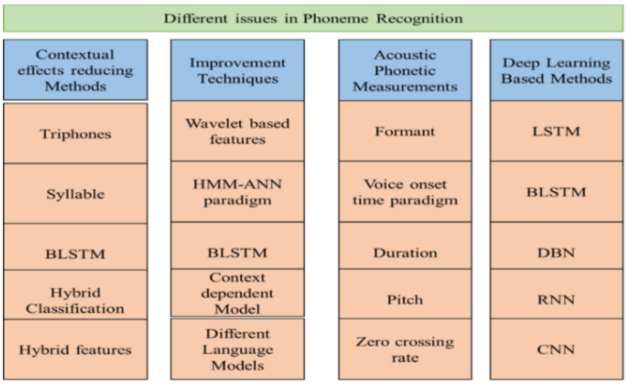
Different acoustic-phonetic measurements were used for phoneme recognition, such as voiced and unvoiced parameters, articulatory features, formant-based measurements, burst and voice onset time, vowel offset and onset points, nasalization, and fricative.

The researchers have suggested different techniques to improve phoneme recognition. Researchers used triphone-based context-dependent phonemes to reduce the contextual effect. To improve Arabic phoneme recognition, helpful statistics based on HMM states for different phonemes were presented.

It was concluded that various measures such as phoneme duration, frequency, and the probability of occurring of different phonemes are helpful statistics to improve phoneme recognition. Further, it was indicated that appropriate states could be designed as per the phoneme. Further researchers also applied wavelet-based features to improve phoneme recognition to address small duration phonemes. Researchers used long-span syllable-based speech recognition to reduce contextual effects. Articulatory features based on place, manner, roundness, frontness, and height were applied to improve phoneme recognition.

The contextual effects in phoneme recognition were addressed using different approaches phoneme recognition was improved by using the HMM-ANN paradigm with the use of contextual information. The contextual effects were examined at the feature level and output level of the multilayer perceptron (MLP). The sub-phonemic groups were addressed at the feature level. The hierarchical estimation of phoneme posterior probabilities was proposed to address contextual information at the output of MLP. For the recognition accuracy measurement, silence was excluded. Bidirectional Long Short-Term Memory (LSTM) network was applied to improve phoneme recognition. Two experiments were performed using the TIMIT speech corpus with unidirectional and bidirectional LSTM networks. The first experiment was conducted for frame-wise phoneme classification. The results indicated that bidirectional LSTM performed better than both unidirectional LSTM and conventional recurrent neural networks (RNNs). It was also stated that RNNs are useful when the span of contextual effects is known. Bidirectional systems present the input into two forward and backward networks better suited for modelling contextual effects. The recent works based on deep neural networks (DNNs) systems proved to give improved speech recognition results. DNN-based systems have the training and testing phases. Though earlier works improved the training phase, this work focused on the test phase. The research work was experimented using a real duration probability distribution for each phoneme using a hidden semi-Markov model (HSMM) instead of geometric distribution of state duration in HMM. Each phoneme was represented by only one state by using phoneme duration in HSMM.

The researchers also examined the performance of a post-processing method that connects the phoneme sequence obtained from the neural network. Bigram language models with MFCCs were used in the experiments. Experiments were conducted on the Persian speech corpus. The results show that the extended Viterbi algorithm on HSMM achieves improvements in phoneme recognition accuracy of 2.68 % and 0.56% over the conventional methods using GMM-HMM and Viterbi on HMM, respectively. It was concluded that RNNs improved speech recognition.



Different issues in Phoneme Recognition

Contextual effects reducing Methods	Improvement Techniques	Acoustic Phonetic Measurements	Deep Learning Based Methods
Triphones	Wavelet based features	Formant	LSTM
Syllable	HMM-ANN paradigm	Voice onset time paradigm	BLSTM
BLSTM	BLSTM	Duration	DBN
Hybrid Classification	Context dependent Model	Pitch	RNN
Hybrid features	Different Language Models	Zero crossing rate	CNN

Figure 6. Phoneme recognition issues related to contextual effects, improvement methods, acoustic-phonetic measurements and deep learning-based methods

It was also indicated that system performance degrades when the testing and training conditions are different. Different databases such as TIMIT, Hindi rhythmic words, Arabic, NTIMIT (telephonic), individual self-created speech corpus, CV/CVC syllable-based corpora, and unique phoneme sets were applied. The speech corpora included polysyllabic words and sentences uttered by male and female speakers with various dialects. TIMIT and NTIMIT (telephonic) were used mostly in the study. The researchers have applied different matrices to evaluate phoneme recognition. Most developers applied the matrices, such as phoneme error rate (PER), phoneme accuracy, and phoneme correctness. Figure 6 shows issues related to contextual effects, improvement methods, acoustic-phonetic measurements and deep learning-based methods.

10. CONCLUSIONS

A comprehensive review of phoneme recognition is presented to explore and understand the different issues. The comprehensive review presented challenges to phoneme recognition. The basic concepts of phoneme recognition were discussed to understand the study. Various essential matters such as phoneme structure, phoneme recognition, vowel recognition, consonant recognition, acoustic-phonetic cues, contextual effects, feature extraction methods, classification methods, and phoneme recognition improvement techniques and performance metrics were covered. The acoustic-phonetic cues for speech recognition such as voiced and unvoiced, articulatory features, formants-based measurement, burst and voice onset time, vowel onset and offset points, nasalization, and fricative were explored. Different methods to calculate acoustic-phonetic cues with challenges and solutions were provided.

Research outcomes reveal that experiments were conducted using different classification methods based on HMM, ANN, GMM, SVM, and VQ. Various feature extraction techniques were also applied to improve phoneme recognition. Different feature extraction techniques were applied, such as MFCCs, LPCs, wavelet-based methods, and PLPs.

Researchers also worked on subcategories such as vowels and consonants in addition to phonemes. It was also observed that different acoustic-phonetic studies were presented to explore the recognition of stop consonants, vowels, semivowels, nasals, and fricatives. It was observed that phoneme recognition suffers from contextual effects. Researchers applied BLSTM ANN-based methods, context-dependent triphones, and longer acoustic units such as syllable-based speech recognition to address the contextual effects.

The research study shows that the researchers mostly worked on identifying the vowels. The vowel recognition was presented using formant analysis, HMM, ANN, linguistic features, and acoustic-phonetic approaches. DNN techniques, hybrid classification (HMM-ANN), articulatory features, landmark-based recognition, event-based recognition, wavelet-based features, improved and hybrid features were applied for phoneme recognition.

Further research may be continued by exploring more confusion analysis of the phonemes.

REFERENCES

[1] Livescu, K., Fosler-Lussier, E., Metze, F. (2012). Subword modeling for automatic speech recognition: Past, present, and emerging approaches. IEEE Signal Processing Magazine, 29(6): 44-57. https://doi.org/10.1109/MSP.2012.2210952

[2] Bhatt, S., Dev, A., Jain, A. (2018). Hindi speech vowel recognition using hidden Markov model. Proc. 6th Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU 2018), pp. 201-204. https://doi.org/10.21437/SLTU.2018-42

[3] Alsulaiman, M., Mahmood, A., Muhammad, G. (2017). Speaker recognition based on Arabic phonemes. Speech Communication, 86: 42-51. https://doi.org/10.1016/j.specom.2016.11.004

[4] Bhatt, S., Jain, A., Dev, A. (2019). CICD acoustic modeling based on monophone and triphone for HINDI speech recognition. In International Conference on Artificial Intelligence and Speech Technology (AIST2019).

[5] Mikolov, T., Zweig, G. (2012). Context dependent recurrent neural network language model. In 2012 IEEE Spoken Language Technology Workshop (SLT), Miami, FL, USA, pp. 234-239. https://doi.org/10.1109/SLT.2012.6424228

[6] Tüske, Z., Sundermeyer, M., Schlüter, R., Ney, H. (2012). Context-dependent MLPs for LVCSR: TANDEM, hybrid or both? In Thirteenth Annual Conference of the International Speech Communication Association, pp. 18-21.

[7] Malakar, M., Keskar, R.B. (2021). Progress of machine learning based automatic phoneme recognition and its prospect. Speech Communication, 135: 37-53. https://doi.org/10.1016/J.SPECOM.2021.09.006

[8] Van Santen, J.P. (1992). Contextual effects on vowel duration. Speech Communication, 11(6): 513-546. https://doi.org/10.1016/0167-6393(92)90027-5

[9] Ganapathiraju, A., Goel, V., Picone, J., Corrada, A., Doddington, G., Kirchhoff, K., Ordowski, M., Wheatley, B. (1997). Syllable-A promising recognition unit for LVCSR. In 1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings, Santa Barbara, CA, USA, pp. 207-214. https://doi.org/10.1109/asru.1997.659007

[10] Espy-Wilson, C.Y. (1994). A feature-based semivowel recognition system. The Journal of the Acoustical Society of America, 96(1): 65-72. https://doi.org/10.1121/1.410375

[11] Hataoka, N., Waibel, A.H. (1990). Speaker-independent phoneme recognition on TIMIT database using integrated time-delay neural networks (TDNNs). In 1990 IJCNN International Joint Conference on Neural Networks, San Diego, CA, USA, pp. 57-62. https://doi.org/10.1109/IJCNN.1990.137544

[12] Lee, J.W., Choi, J.Y. (2008). Acoustic-phonetic features for stop consonant place detection in clean and telephone speech. Journal of the Acoustical Society of America, 123(5): 3330. https://doi.org/10.1121/1.2933843

[13] Jassem, W. (2014). Acoustic-phonetic variability of polish vowels. Archives of Acoustics, 17(2): 217-233.

[14] Sarma, B.D., Prasanna, S.M. (2018). Acoustic–phonetic analysis for speech recognition: A review. IETE Technical Review, 35(3): 305-327. https://doi.org/10.1080/02564602.2017.1293570

[15] Michalek, J., Vaněk, J. (2018). A survey of recent DNN architectures on the TIMIT phone recognition task. In Text, Speech, and Dialogue: 21st International Conference, TSD 2018, Brno, Czech Republic, September 11-14, 2018, Proceedings 21, pp. 436-444. Springer International Publishing. https://doi.org/10.1007/978-3-030-00794-2_47

[16] AlDahri, S.S., Alotaibi, Y.A. (2010). A crosslanguage survey of VOT values for stops (/d/,/t/). In 2010 IEEE International Conference on Intelligent Computing and Intelligent Systems, Xiamen, China, 3: 334-338.

https://doi.org/10.1109/ICICISYS.2010.5658744

[17] Nirmala, S.R., Upashana, G. (2017). A review on landmark detection methodologies of stop consonants. Advances in Computational Research, Bioinfo Publication, 8(1): 316-320.

[18] Bhatt, S., Dev, A., Jain, A. (2022). Hindi phoneme recognition - A review. In: Dev, A., Agrawal, S.S., Sharma, A. (eds) Artificial Intelligence and Speech Technology. AIST 2021. Communications in Computer and Information Science, vol 1546. Springer, Cham. https://doi.org/10.1007/978-3-030-95711-7_4

[19] Kshirsagar, A., Dighe, A., Nagar, K., Patidar, M. (2012). Comparative study of phoneme recognition techniques. In 2012 Third International Conference on Computer and Communication Technology, Allahabad, India, pp. 98-103. https://doi.org/10.1109/ICCCT.2012.28

[20] Samad, A., Rehman, A.U., Ali, S.A. (2019). Performance evaluation of learning classifiers of children emotions using feature combinations in the presence of noise. Engineering, Technology & Applied Science Research, 9(6): 5088-5092.

[21] Ibrahim, A.B., Seddiq, Y.M., Meftah, A.H., Alghamdi, M., Selouani, S.A., Qamhan, M.A., Alotaibi, Y.A., Alshebeili, S.A. (2020). Optimizing arabic speech distinctive phonetic features and phoneme recognition using genetic algorithm. IEEE Access, 8: 200395-200411. https://doi.org/10.1109/ACCESS.2020.3034762

[22] Veisi, H., Haji Mani, A. (2020). Persian speech recognition using deep learning. International Journal of Speech Technology, 23: 893-905. https://doi.org/10.1007/s10772-020-09768-x

[23] Laleye, F.A., Ezin, E.C., Motamed, C. (2016). Automatic fongbe phoneme recognition from spoken speech signal. In Proceedings of the 13th International Conference on Informatics in Control, Automation and Robotics (ICINCO 2016), pp. 102-109. https://doi.org/10.5220/0006004101020109

[24] Qamhan, M.A., Alotaibi, Y.A., Seddiq, Y.M., Meftah, A.H., Selouani, S.A. (2021). Sequence-to-sequence acoustic-to-phonetic conversion using spectrograms and deep learning. IEEE Access, 9: 80209-80220. https://doi.org/10.1109/ACCESS.2021.3083972

[25] Paul, B., Phadikar, S. (2023). A novel pre-processing technique of amplitude interpolation for enhancing the classification accuracy of bengali phonemes. Multimedia Tools and Applications, 82(5): 7735-7755. https://doi.org/10.1007/S11042-022-13594-5

[26] Karpagavalli, S., Chandra, E. (2016). A review on sub-word unit modeling in automatic speech recognition. IOSR Journal of VLSI and Signal Processing, 6(6): 2319-4197.

[27] Killer, M., Stüker, S., Schultz, T. (2003). Grapheme based speech recognition. Master Thesis, The Interactive Systems Laboratory.

[28] Duncan, L.G. (2018). Language and reading: The role of morpheme and phoneme awareness. Current Developmental Disorders Reports, 5: 226-234. https://doi.org/10.1007/s40474-018-0153-2

[29] Ferrer, L., Lei, Y., McLaren, M., Scheffer, N. (2014). Spoken language recognition based on senone posteriors. In Fifteenth Annual Conference of the International Speech Communication Association, pp. 2150-2154.

[30] Yoshida, K., Watanabe, T. (1989). Large vocabulary word recognition based on demi-syllable hidden Markov model using small amount of training data. In International Conference on Acoustics, Speech, and Signal Processing, Glasgow, UK, pp. 1-4. https://doi.org/10.1109/ICASSP.1989.266348

[31] Dahl, G.E., Yu, D., Deng, L., Acero, A. (2011). Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. IEEE Transactions on Audio, Speech, and Language Processing, 20(1): 30-42. https://doi.org/10.1109/TASL.2011.2134090

[32] Pfeifer, V., Balik, M. (2011). Comparison of current frame-based phoneme classifiers. Advances in Electrical and Electronic Engineering, 9(5): 243-250. https://doi.org/10.15598/aeee.v9i5.545

[33] Salvi, G. (1999). Developing acoustic models for automatic speech recognition in Swedish. European Student Journal of Language and Speech, Stockholm, Sweden.

[34] Deller Jr, J.R., Hansen, J. (2004). Methods, models, and algorithms for modern speech processing. The Electrical Engineering Handbook, 861-889.

[35] Brietzmann, A. (1982). Pragmatics in speech understanding-revisited. In Coling 1982: Proceedings of the Ninth International Conference on Computational Linguistics, pp. 49-54.

[36] Young, S.R. (1990). Use of dialogue, pragmatics and sematics to enhance speech recognition. Speech Communication, 9(5-6): 551-564. https://doi.org/10.1016/0167-6393(90)90030-D

[37] MY, S.A., Yaacob, S., Paulraj, M.P. (2009). Vowel recognition using first formant feature. International Journal of Advancements in Computing Technology, 1(1): 24-31.

[38] Pruthi, T., Espy-Wilson, C. (2003). Automatic classification of nasals and semivowels. In ICPhS 2003-15th International Congress of Phonetic Sciences, pp. 3061-3064.

[39] Baltazani, M. (2007). Prosodic rhythm and the status of vowel reduction in Greek. Selected Papers on Theoretical and Applied Linguistics, 17(1): 31-43. https://doi.org/10.26262/ISTAL.V17I1.5539

[40] Whalen, D.H., Chen, W.R. (2019). Variability and central tendencies in speech production. Frontiers in Communication, 4: 49. https://doi.org/10.3389/fcomm.2019.00049

[41] Cavalcanti, J.C., Eriksson, A., Barbosa, P.A. (2021). Acoustic analysis of vowel formant frequencies in genetically-related and non-genetically related speakers with implications for forensic speaker comparison. Plos One, 16(2): e0246645. https://doi.org/10.1371/JOURNAL.PONE.0246645

[42] Manjunath, K.E., Sreenivasa Rao, K. (2016). Articulatory and excitation source features for speech recognition in read, extempore and conversation modes. International Journal of Speech Technology, 19: 121-134. https://doi.org/10.1007/s10772-015-9329-x

[43] Welling, L., Ney, H. (1998). Formant estimation for speech recognition. IEEE Transactions on Speech and Audio Processing, 6(1): 36-48. https://doi.org/10.1109/89.650308

[44] Abramson, A.S., Whalen, D.H. (2017). Voice Onset Time (VOT) at 50: Theoretical and practical issues in measuring voicing distinctions. Journal of Phonetics, 63: 75-86. https://doi.org/10.1016/j.wocn.2017.05.002

[45] Shrem, Y., Goldrick, M., Keshet, J. (2019). Dr. VOT: Measuring positive and negative voice onset time in the wild. arXiv preprint arXiv:1910.13255. https://doi.org/10.48550/arXiv.1910.13255

[46] Feng, G., Castelli, E. (1996). Some acoustic features of nasal and nasalized vowels: A target for vowel nasalization. The Journal of the Acoustical Society of America, 99(6): 3694-3706. https://doi.org/10.1121/1.414967

[47] Chen, M.Y. (1997). Acoustic correlates of English and French nasalized vowels. The Journal of the Acoustical Society of America, 102(4): 2360-2370. https://doi.org/10.1121/1.419620

[48] Bhatt, S., Dev, A., Jain, A. (2020). Confusion analysis in phoneme based speech recognition in Hindi. Journal of Ambient Intelligence and Humanized Computing, 11: 4213-4238. https://doi.org/10.1007/s12652-020-01703-x

[49] Lee, K.F. (1990). Context-independent phonetic hidden Markov models for speaker-independent continuous speech recognition. IEEE Transactions on Acoustics, Speech, and Signal Processing, 38(4): 599-609. https://doi.org/10.1109/29.52701

[50] Graves, A., Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Networks, 18(5-6): 602-610. https://doi.org/10.1016/j.neunet.2005.06.042

[51] Pinto, J., Yegnanarayana, B., Hermansky, H., Magimai-Doss, M. (2008). Exploiting contextual information for improved phoneme recognition. In 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, Las Vegas, NV, USA, pp. 4449-4452. https://doi.org/10.1109/ICASSP.2008.4518643

[52] Amami, R., Ellouze, N. (2015). Study of phonemes confusions in hierarchical automatic phoneme recognition system. arXiv preprint arXiv:1508.01718. https://doi.org/10.48550/arXiv.1508.01718

[53] Lee, C.H., Rabiner, L.R., Pieraccini, R., Wilpon, J.G. (1990). Acoustic modeling of subword units for speech recognition. In International Conference on Acoustics, Speech, and Signal Processing, Albuquerque, NM, USA, pp. 721-724. https://doi.org/10.1109/icassp.1990.115885

[54] Khelifa, M.O., ElHadj, Y.O., Abdellah, Y., Belkasmi, M. (2017). Helpful statistics in recognizing basic Arabic phonemes. International Journal of Advanced Computer Science and Applications (IJACSA), 8(2): 238-244. https://doi.org/10.14569/ijacsa.2017.080231

[55] Farooq, O., Datta, S., Shrotriya, M. C. (2010). Wavelet sub-band based temporal features for robust Hindi phoneme recognition. International Journal of Wavelets, Multiresolution and Information Processing, 8(6): 847-859. https://doi.org/10.1142/S0219691310003845

[56] Ganapathiraju, A., Hamaker, J., Picone, J., Ordowski, M., Doddington, G.R. (2001). Syllable-based large vocabulary continuous speech recognition. IEEE Transactions on Speech and Audio Processing, 9(4): 358-366. https://doi.org/10.1109/89.917681

[57] Asadolahzade Kermanshahi, M., Homayounpour, M.M. (2019). Improving phoneme sequence recognition using phoneme duration information in DNN-HSMM. Journal of AI and Data Mining, 7(1): 137-147. https://doi.org/10.22044/JADM.2018.6136.1725

[58] Moses, D.A., Mesgarani, N., Leonard, M.K., Chang, E.F. (2016). Neural speech recognition: Continuous phoneme decoding using spatiotemporal representations of human cortical activity. Journal of Neural Engineering, 13(5): 056004. https://doi.org/10.1088/1741-2560/13/5/056004

[59] Vasquez, D., Gruhn, R., Minker, W. (2012). Hierarchical Neural Network Structures for Phoneme Recognition. Springer Science & Business Media.

[60] Gales, M., Young, S. (2008). The application of hidden Markov models in speech recognition. Foundations and Trends® in Signal Processing, 1(3): 195-304. http://dx.doi.org/10.1561/2000000004

[61] Schwarz, P. (2008). Phoneme recognition based on long temporal context. Doctoral dissertation, Ph. D. dissertation, Faculty of Information Technology, Brno University of Technology.

[62] Koolagudi, S.G., Thakur, S.N., Barthwal, A., Singh, M.K., Rawat, R., Sreenivasa Rao, K. (2012). Vowel recognition from telephonic speech using MFCCs and Gaussian mixture models. In: Mathew, J., Patra, P., Pradhan, D.K., Kuttyamma, A.J. (eds) Eco-friendly Computing and Communication Systems. ICECCS 2012. Communications in Computer and Information Science, vol 305. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-32112-2_21

[63] Biswas, A., Sahu, P.K., Bhowmick, A., Chandra, M. (2014). Hindi vowel classification using GFCC and formant analysis in sensor mismatch condition. WSEAS Transactions on Systems, 13: 130-143.

[64] Khwaja, M.K., Vikash, P., Arulmozhivarman, P., Lui, S. (2016). Robust phoneme classification for automatic speech recognition using hybrid features and an amalgamated learning model. International Journal of Speech Technology, 19: 895-905. https://doi.org/10.1007/s10772-016-9377-x