International Journal of Computer Applications (0975 - 8887)
Volume 175 - No.30, O 2020

A Revisit to Speech Processing and Analysis

Aniruddha Mohanty
CHRIST (DEEMED TO BE UNIVERSITY)
Bangalore, 560029
India

Ravindranath C. Cherukuri
CHRIST (DEEMED TO BE UNIVERSITY)
Bangalore, 560029
India

ABSTRACT

Speech recognition is an active area in signal processing.Various researchers have been invested different concepts in speech recognition system as part of feature extraction techniques, speech classifiers, statistical analysis, encompassing mathematical models, signal processing and transformations, database and performance evaluation. In the current era, multi speaker analysis is the newly focused area in speech processing and analysis. It includes audio segmentation, extraction of relevant features, classification of features, template generation and training. Also, other techniques like Bank-of-filters, Linear Predictive Coding Model, Vector Quantization, Hidden Markov Model and Gaussian Mixture Model to get better result. In this paper, various approaches have been analyzed based on acoustic and articular features focusing on Human Auditory System (HAS). Even focusing on the cross functional approach by using machine learning, artificial intelligence-based techniques and neural networks.

General Terms

Automatic speech recognition, modelling and matching techniques

Keywords

Automatic speech recognition, feature extraction, dimension reduction, modeling and matching techniques

1. INTRODUCTION

Communication is a powerful mechanism for sharing the information. In human-machine interaction helps to identify the design, evaluation and implementation of interactive computing systems for human. The aim of the speech recognition task is to address the speaker by dividing the speech sample into small segments. There are so many languages like regional, national and international languages that have been spoken around the globe. It is needed to recognize someone by listening the speech if the language is understood. This recognition system has been implemented in different areas such as personal computers, mobile phones, security systems, healthcare, robotics, military, education, dictation and many more.

Currently, the Automatic Speech Recognition (ASR) system uses acoustic and articulatory features from speakers in noisy environments like car environment, reverberant environment and other vehicular environment. Before implementing the model to recognize the speech, samples need to be preprocessed to desire scales and overlapped with each other to get the smooth results.

2. CLASSIFICATION OF SPEECH RECOGNITION SYSTEM

The speech recognition system is based on utterance, vocabulary size and speaker[1]

2.1 Classification Based on Utterances

Speech recognition system can be segregated into different classes by describing what types of utterances in the speech[2]. These are classified as follows. Isolated word recognizers usually require each utterance to have quiet (lack of an audio signal) on both sides of the sample window. It accepts single word or single utterance at a time. The systems having ”Listen/Not-Listen” states, where they require the speaker to wait between utterances [2] [3]. Connected word systems are similar to isolated words, but allows separate utterances to be run-together with a minimal pause between them [2] [4]. Continuous speech is the speech having words are connected (words are not separated by breathing space). It is difficult to find the start and end points of the words [1] [5].Spontaneous Speech is generated by disrupted air flow generated in the vocal tract, nonlinear neuromuscular processes may take place at the larynx and the level of vocal cords [1] [6].

2.2 Classification Based on Vocabulary Size

The computational complexity, processing requirement and the accuracy of the speech recognition system depends on the vocabulary size of the speech [7]. As the vocabulary size increases, the task of recognition system become quite tedious. ASR system is classified based on the vocabulary as follows [1].
i. Small Vocabulary -10 Words
ii. Medium Vocabulary - 100 Words
iii. Large Vocabulary - 1000 Words
iv. Very-Large Vocabulary -10000 Words
v. Out-of-Vocabulary - Mapping a word from the vocabulary into the unknown words.

Speaker Independent Systems do not require a user to train the system i.e. they are developed to operate for any speaker [8]. Speaker Dependent Systems need user to train the system according to users voice. Train data and the test data are from the same set of speakers [8].

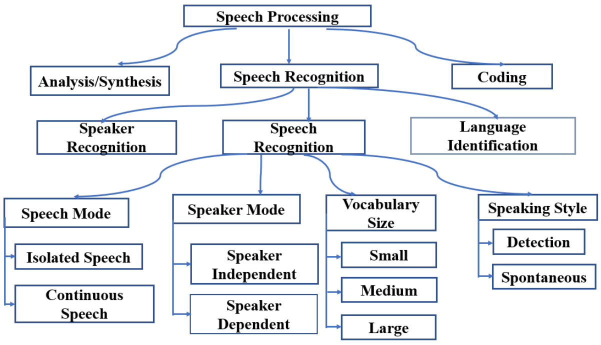

Fig. 1. Speech Processing Classification.

3. SPEECH ANALYSIS

Speech analysis are often evaluated in various conditions including environment, microphone, data simulation mismatch. Generally, four steps should be followed in order to design the speech recognition system.
i. Analysis
ii. Feature Extraction
iii. Modelling
iv. Testing

3.1 Analysis

Speech is used as a tool for manipulating electronic devices. In Human-machine interaction, a lot of desktop and web based services are used to provide the functionality of automatic dictation, voice search etc.[9].

3.1.1 Sampling and Quantization:. Most speech signals are nonstationary processes with multiple components that may vary in time and frequency [10]. So, speech signal should be digitized the continuous-time signal by the help of sampling and quantization. The continuous signal x(t) is sampled to give x(n), which yields $x_Q(n)$ after quantization [11].

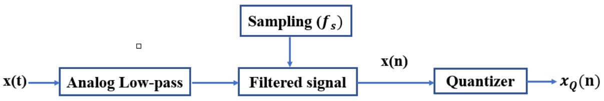

Fig. 2. Analog to Digital Conversion.

3.1.2 Preprocessing:. Preprocessing is the very first step after collecting the speech data, used to extract local spectral information. It follows pre-emphasis step to amplify the high frequency components. After that segmentation steps would be followed [12]. Speech Framing is also called speech segmentation. In this process, the continuous speech is segregated into fixed length segments for accurate decision. Partition of the speech samples into short frames makes the signal stationary or quasistationary states, which helps to obtain local features [13]. Again, each current frame has 30% to 50% overlap with the previous frame to ward off the information loss [12]. After the framing of the speech signal, the frames have been applied a window function which is used to split the input signal to temporal segment and the borders of the segments are visible as discontinued. Several windowing techniques like Rectangular, Hamming, Hann, Blackman are used to design the ASR system [14]. The window function for window size N and frame w(x) is expressed as

$$W(X) = 0.54 - 0.46cos\frac{2\pi n}{N-1}, \qquad 0 \le X \le N-1 \qquad (1)$$

Normalization is applied to the speech samples which is used to reduce speaker and recording variability without losing the strength of the features [15]. If mean μ and standard deviation σ of the data, standardization is expressed as

$$X_s = \frac{x-\mu}{\sigma} \qquad (2)$$

To get high speech recognition rate, Noise Reduction techniques must be used to reduce the environmental or background noises [16]. Minimum mean square error (MMSE) and log-spectral amplitude MMSE (LogMMSE) estimators are mostly used for noise reduction [13] [17]. Even filtering techniques like spectral subtraction, spectrum reconstruction and regression model are also used for noise Reduction [18].

3.2 Feature Extraction

In Speech analysis, Feature extraction helps the feature extractor to keep relevant information and discard the irrelevant information which helps the formulation of the Speaker Identification and verification system [1]. Speech is a continuous signal of varying length. Below global or local features can be extracted to analyze the ASR system [13].
i. Prosodic Features
ii. Spectral Features
iii. Voice Quality Features
iv. Teager Energy Operator (TEO) Based Features

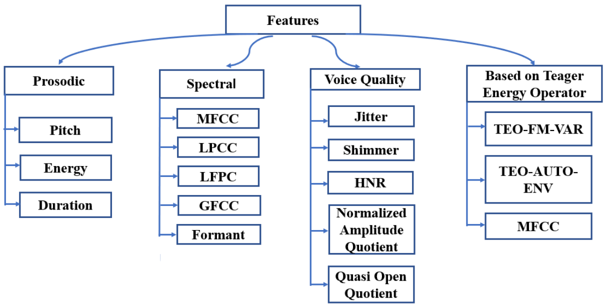

Fig. 3. Speech Features.

3.2.1 Prosodic Features. There are several prosodic features that influence speech e.g. Pitch, Formant and Energy or Loudness etc. [19]. Based on segmentation, continuous speech is divided into syllable-like units, sentences/phrases or even fixed-interval segments [20]. Prosodic features are divided into
i. Inflections or start/end voicing
ii. Information from F0 and energy contour
iii. Detection of voice region
Pitch is a feature defined as the number of periods of vocal folds vibration per second. The pitch is calculated frame by frame based on the robust algorithm for pitch tracking [21]. Formant is basically referred as the acoustic resonance of human vocal tract. The concentration of acoustic energy is also detected within a particular range of frequency. This can also be measured by

frequency peak in the spectrum [22]. The amplitude of unvoiced segments is generally much lower than the amplitude of voiced segments. The signal does not contain unvoiced segments, its Short Time Energy (STE) is usually bigger. For a discrete signal S(n), the energy is given by

$$E_s = \sum_{n=\infty}^{\infty} |S(n)|^2 \tag{3}$$

The power of discrete-time signal S(n) is given by [23].

$$P_s = \lim_{N=\infty} \frac{1}{2N+1} \sum_{n=N}^{N} |S(n)|^2 \tag{4}$$

In general, signals can be classified into three types: an energy signal, which has a non-zero and finite energy, a power signal and the third type is neither. Zero-crossing rate is one parameter which is used to find the voiced and unvoiced portion of the speech. At first the speech sample is divided into frames, then zero-crossing is calculated form each frame to detect the voiced and unvoiced portion of the speech [24]. Auto-correlation is a measure of self-similarity of a signal in time domain also the similarity between the signal and its delayed version. The auto-correlation value of positive 1 (+1) represents strong positive association, negative 1 (-1) represents a negative association and 0 shows no association. Auto-correlation function is used to determine the periodicity present in a signal and used to estimate the pitch (fundamental frequency) of a signal [25]. This also used for Noise-Robust Speaker Recognition [26]. The Wavelet Transform gives a similarity with how human ear processes sound. Therefore, it is suitable for speech processing. It is of two types Discrete Wavelet Transform (DWT) and Wavelet Packet Decomposition (WPD) [27]. DWT operates in discrete steps over signal and provide sufficient information with varying window size, being wide of low frequency and narrow for high frequency [28]. WPD to obtain high frequency and low frequency set of coefficients which allows better time-frequency localization of signals [29].

3.2.2 Spectral Features. Speech signals having segmental spectral features contain enough information, which helps for speech recognition. The features such as Linear Predictive Cepstral coefficients (LPCC) and Mel-Frequency Cepstral Coefficients (MFCC) [30] [31] are used frequently. Spectral features are obtained by transforming the time domain signal into the frequency domain signal by using the Fourier transform by segmenting the speech samples to 20 to 30 milliseconds [13]. Mel Frequency Cepstral Coefficients (MFCC) represents short-time power spectrum of an audio clip based on the discrete cosine transform of log power spectrum on a nonlinear equally spaced mel scale which mimics the human auditory system [25]. Linear Predictive Coding (LPC) draws out parameters of speech like spectra and pitch formants. It also equivalents to the resonance structure of human vocal which reduce the square difference between original and estimated speech signal at a finite time. Gammatone Frequency Cepstral Coefficients (GFCC) is obtained by windowing and Fourier transforms similar to MFCC then filtered using the Gammatone filters. This also gives the Time-Frequency representation [32]. A spectrogram represents the strength or loudness of a signal over time at different frequencies in a particular waveform. These are used to verify the frequencies in continuous signals and represented by a graph with two geometric dimensions in which time is viewed on the horizontal axis, while the vertical axis identified by frequency and the intensity or color of each point in the image corresponds to amplitude of particular frequency at particular time. Short Term Fourier Transform (STFT) is usually applied to the speech sample to generate spectrogram from the time signal. Using Fast Fourier Transform (FFT) for generating the spectrogram is a digital process [33].

3.2.3 Voice Quality Features. Voice quality is determined by the physical properties of the vocal tract. Spontaneous changes may produce a speech signal that might differentiate speech by using properties such as jitter, shimmer, and Harmonics to Noise Ratio (HNR) [13].
Jitter is the variability of fundamental frequency between successive vibratory cycles, while shimmer is the variable of the amplitude. Jitter is a measure of frequency instability, whereas shimmer is the amplitude instability.
The voice qualities are grouped into the following categories [34].
i. Voice Level: Signal amplitude, energy and duration.
ii. Voice pitch
iii. Phase, phenome, word and feature boundaries

3.2.4 Teager energy operator-based features. According to Teager, speech is formed by a non-linear vortex-airflow interaction in the human vocal system. An exciting situation affects the muscle tension of the speaker that results in an alteration of the airflow during the production of the sound. The operator developed by Teager to measure the energy from a speech by this non-linear process was documented by Kaiser as follows where $\psi[]$ is Teager Energy Operator and x(n) is the sampled speech signal or discrete signal [13].

$$\psi[X(n)] = x^2(n) - x(n+1)x(n-1) \tag{5}$$

The TEO for a continuous time signal is defined as [35].

$$\psi(x(n)) = \frac{\partial x(t)^2}{\partial t} - x(t)(\frac{\partial^2 x(t)}{\partial t^2}) \tag{6}$$

3.2.5 Entropy. Entropy measures the amount of information content in a signal. By the help of Boltzmans formula is determined how much a signal could be compressed, the more information content in a signal the less it could be compressed.

$$H(n) = -\sum_{i=1}^{M} P[X = x_i] log([P(X = x_i)]) \tag{7}$$

where M is the number of possible values of discrete random variable X (i.e. the signal), x_i is the i^{th} of these values. Spectral entropy has been determined for each frame but not decomposed in frequency bands [36].

4. FEATURE SELECTION AND DIMENSION REDUCTION

Feature selection methods aim to select a subset of the original high-dimensional features based on some performance criterion. Therefore, it can be preserved the semantics of the original features and produce dimensionally reduced results that are more interpretable for domain experts [37]. Mostly the feature selection algorithms are grouped into following categories.
i. Similarity based feature selection
ii. Information theoretical based feature selection
iii. Statistical based feature selection

In similarity based feature selection category, data similarity has been preserved. By Laplacian Score method constructs a nearest neighbor graph for dataset samples and use the Laplacian matrix to calculate a score for each feature which shows the importance of the feature in preserving the similarity and locality of dataset samples [38]. Information theoretical based feature selection method uses the label data to consider the relevance and redundancy of features by maximal statistical dependency criterion based on mutual information [39]. Statistical based feature selection, this category based on various types of statistical procedures which evaluates individually. Variance-score, t-score and chi-score are well-known representatives of this family [38]. Dimension Reduction is to transform of data from a high-dimensional space into a low-dimensional space by which some meaningful properties of the original data can be retained and also reduce the amount of overtraining. Commonly used techniques are Principal Component Analysis (PCA) [40] and even Linear Discriminant Analysis (LDA). PCA is an unsupervised learning dimensionality reduction method by mapping high dimensionality feature set to low dimensionality space, in which feature set is decomposed by the covariance matrix to obtain the principal components and weights of feature set. It utilizes the variance of data projection to assess the amount of feature representation information. PCA is intended to select first k-dimension features with the largest variance to ensure the data after projection satisfies the variance maximization [41] [42]. LDA is a supervised dimensionality reduction method used as a dimensionality reduction and feature extraction tool in pattern recognition. This measures the information with the difference of labels and categories. LDA can be class-dependent or class-independent, based upon maximization of the ratio between class variance to within class variance or maximization of the ratio of overall variance to within class variance respectively [43] [44]. The Joint principal component and discriminant analysis, is used to transform the initial feature set into a new subspace and also extract the discriminant information classification task [45].

5. MODELLING

The main purpose of modelling technique is to generate speaker models using speaker specific feature vector. The modeling techniques are used for speaker recognition and speaker identification. The Acoustic-Phonetic approach is based on acoustic phonetics and postulates particularly the acoustic phonetic search. Using International Phonetic Alphabet (IPA) methods, Similarities for probabilities of content dependant acoustic model for new language can be found out. Pattern Recognition approach involves two steps, pattern training and pattern comparison. This approach used to decide whether the given speech segment belongs to voiced, unvoiced or silence and also can be applied to a sound, a word, a phrase [46]. Dynamic time warping (DTW) is the simplest way to recognize Variations of word structure (VoWS), Normalized Phoneme Distances Thresholding (NPDT), Furthest Segment Search (FSS) and Normalized Furthest Segment Search (NFSS) [47]. The artificial intelligence approach is a hybrid of the acoustic phonetic approach and pattern recognition approach, is used to mechanize the recognition procedure according to the way a human being applies its intelligence. Support Vector Machine (SVMs) is the binary classifier consists of linear and nonlinear separating hyperplanes for information categorization. This model consists of classification of static information vectors only. But this technique cannot be classifying the dynamic data. Also operates the complexity of the model by controlling the vector quantization dimensions of the model [48].

5.1 MATCHING TECHNIQUES

In the next step, it needs to match a detected word from a known word. This can achieve by whole word matching and sub word matching [49]. Whole Word Matching framework compares the incoming digital-audio signal against a prerecorded template of the words having connected word string spoken frequently [50]. Sub Word Matching framework looks for sub-words usually phonemes and then performs further pattern recognition on those [51].

5.2 PERFORMANCE OF SYSTEMS

A speech recognition engine recognizes all words uttered by a human but, practically the performance of a speech recognition engine depends on a number of factors. Vocabularies, multiple users and noisy environments [52]. Word Error Rate (WER) is the common metric of the accuracy of a speech recognition system. WER is derived from the Levenshtein distance, working at the word level instead of the phoneme level [53]. In Command Success Rate, system stores data vocabularies and the vocabulary contain the ten digits from 0 to 9 mapped with ten command words like enter, erase, go, help, no, rubout, repeat, stop, start, and yes. Based on the commands, the training and the testing data set can be identified [54].

6. CHALLENGES

Although there are many amendments in speech recognition system, still several impediments that need to be removed to make a successful recognition system. First problem is getting or creation of the proper data set as per the required task. Most of the data is embedded with noise and other environmental distractions. Some lacunae are there to identify the speech recognition system in unsupervised manner so that the performance of the computer assistants can be improved. Again, the design of model for speech recognition system should focus on individuals own way of speaking, which depends upon various factors that may include the dialect and accent of the speaker as well as the socioeconomic background of the speaker. Among various speech processing problems, Automatic Speech Recognition (ASR) for converting recorded speech automatically to text is one of the most challenging tasks.

7. CONCLUSIONS

In this survey paper, multiple research work in speech recognition with feature extraction, Speech recognition model design have been described. Basically, preprocessing is required to remove the noise from the speech sample and make the sample processable. Then features are needed to analyse the speech signal differently as part of the various domains. After that the features are processed in a designed speech system to get an intended performance. The focus should be on the previous studies and the used techniques to know the benefits of the designed system and their performances. This also covers accent, speaking style, speaker physiology, age, emotions. General methods for diagnosing weaknesses in speech recognition approaches are also highlighted. Finally, the paper proposed an overview of general and specific techniques for better handling of variation sources in Automatic Speech Recognition.

8. REFERENCES

[1] Praphulla A Sawakare, Ratndeep R Deshmukh, and Pukhraj P Shrishrimal. Speech recognition techniques: A review. *International Journal of Scientific & Engineering Research*, 6(8):1693–1698, 2015.

[2] MA Anusuya and Shriniwas K Katti. Speech recognition by machine, a review. *arXiv preprint arXiv:1001.2267*, 2010.

[3] UG Patil, SD Shirbahadurkar, and AN Paithane. Automatic speech recognition of isolated words in hindi language using mfcc. In *2016 International Conference on Computing, Analytics and Security Trends (CAST)*, pages 433–438. IEEE, 2016.

[4] Shobha Bhatt, Amita Dev, and Anurag Jain. Confusion analysis in phoneme based speech recognition in hindi. *Journal of Ambient Intelligence and Humanized Computing*, pages 1–26, 2020.

[5] PS Praveen Kumar and HS Jayanna. Performance analysis of hybrid automatic continuous speech recognition framework for kannada dialect. In *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pages 1–6. IEEE, 2019.

[6] Mahda Nasrolahzadeh, Zeynab Mohammadpoory, and Javad Haddadnia. Higher-order spectral analysis of spontaneous speech signals in alzheimers disease. *Cognitive neurodynamics*, 12(6):583–596, 2018.

[7] Muhammad Atif Imtiaz and Gulistan Raja. Isolated word automatic speech recognition (asr) system using mfcc, dtw & knn. In *2016 Asia Pacific Conference on Multimedia and Broadcasting (APMediaCast)*, pages 106–110. IEEE, 2016.

[8] Ratnadeep Deshmukh and Abdulmalik Alasadi. Automatic speech recognition techniques: A review, 2018.

[9] Aidar Khusainov and Alfira Khusainova. Speech analysis and synthesis systems for the tatar language. In *2016 IEEE Artificial Intelligence and Natural Language Conference (AINL)*, pages 1–6. IEEE, 2016.

[10] Vinod Chandran and Boualem Boashash. Time-frequency methods in radar, sonar, and acoustics. *Time-frequency signal analysis and processing (Second Edition): A comprehensive reference*, pages 793–856, 2016.

[11] Udo Zölzer. *Digital audio signal processing*, volume 9. Wiley Online Library, 2008.

[12] Jibin Wu, Yansong Chua, and Haizhou Li. A biologically plausible speech recognition framework based on spiking neural networks. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2018.

[13] Mehmet Berkehan Akçay and Kaya Oğuz. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Communication*, 116:56–76, 2020.

[14] Zulfiqar Ali, M Shamim Hossain, Ghulam Muhammad, and Arun Kumar Sangaiah. An intelligent healthcare system for detection and classification to discriminate vocal fold disorders. *Future Generation Computer Systems*, 85:19–28, 2018.

[15] Ronald Böck, Olga Egorow, Ingo Siegert, and Andreas Wendemuth. Comparative study on normalisation in emotion recognition from speech. In *International Conference on Intelligent Human Computer Interaction*, pages 189–201. Springer, 2017.

[16] Win Lai Lai Phyu and Win Pa Pa. Building speaker identification dataset for noisy conditions. In *2020 IEEE Conference on Computer Applications (ICCA)*, pages 1–6. IEEE, 2020.

[17] Chengli Sun, Qi Zhu, and Minghua Wan. A novel speech enhancement method based on constrained low-rank and sparse matrix decomposition. *Speech Communication*, 60:44–55, 2014.

[18] Bo Zheng, Jinsong Hu, Ge Zhang, Yuling Wu, and Jianshuang Deng. Analysis of noise reduction techniques in speech recognition. In *2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, volume 1, pages 928–933. IEEE, 2020.

[19] Atreyee Khan and Uttam Kumar Roy. Emotion recognition using prosodie and spectral features of speech and naïve bayes classifier. In *2017 international conference on wireless communications, signal processing and networking (WiSPNET)*, pages 1017–1021. IEEE, 2017.

[20] Leena Mary. Extraction and representation of prosody for speaker, language, emotion, and speech recognition. In *Extraction of Prosody for Automatic Speaker, Language, Emotion and Speech Recognition*, pages 23–43. Springer, 2019.

[21] Safa Chebbi and Sofia Ben Jebara. On the use of pitch-based features for fear emotion detection from speech. In *2018 4th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*, pages 1–6. IEEE, 2018.

[22] Kasiprasad Mannepalli, Panyam Narahari Sastry, and Maloji Suman. Analysis of emotion recognition system for telugu using prosodic and formant features. In *Speech and Language Processing for Human-Machine Communications*, pages 137–144. Springer, 2018.

[23] Abdullah I Al-Shoshan. Speech and music classification and separation: a review. *Journal of King Saud University-Engineering Sciences*, 19(1):95–132, 2006.

[24] Priyanka Gupta and S Sengupta. Voiced/unvoiced decision with a comparative study of two pitch detection techniques. 2018.

[25] Garima Sharma, Kartikeyan Umapathy, and Sridhar Krishnan. Trends in audio signal feature extraction methods. *Applied Acoustics*, 158:107020, 2020.

[26] Paavo Alku and Rahim Saeidi. The linear predictive modeling of speech from higher-lag autocorrelation coefficients applied to noise-robust speaker recognition. *IEEE/acm transactions on audio, speech, and language processing*, 25(8):1606–1617, 2017.

[27] Mohammed Arif Mazumder and Rosalina Abdul Salam. Feature extraction techniques for speech processing: A review. 2019.

[28] Khushboo S Desai and Heta Pujara. Speaker recognition from the mimicked speech: A review. In *2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, pages 2254–2258. IEEE, 2016.

[29] Yong Yin, Yu Bai, Fei Ge, Huichun Yu, and Yunhong Liu. Long-term robust identification potential of a wavelet packet decomposition based recursive drift correction of e-nose data for chinese spirits. *Measurement*, 139:284–292, 2019.

[30] Yu Zhou, Yanqing Sun, Jianping Zhang, and Yonghong Yan. Speech emotion recognition using both spectral and prosodic

features. In *2009 International Conference on Information Engineering and Computer Science*, pages 1–4. IEEE, 2009.

[31] Agustinus Bimo Gumelar, Afid Kurniawan, Adri Gabriel Sooai, Mauridhi Hery Purnomo, Eko Mulyanto Yuniarno, Indar Sugiarto, Agung Widodo, Andreas Agung Kristanto, and Tresna Maulana Fahrudin. Human voice emotion identification using prosodic and spectral feature extraction based on deep neural networks. In *2019 IEEE 7th International Conference on Serious Games and Applications for Health (SeGAH)*, pages 1–8. IEEE, 2019.

[32] Mohit Dua, Rajesh Kumar Aggarwal, and Mantosh Biswas. Gfcc based discriminatively trained noise robust continuous asr system for hindi language. *Journal of Ambient Intelligence and Humanized Computing*, 10(6):2301–2314, 2019.

[33] Abdul Malik Badshah, Nasir Rahim, Noor Ullah, Jamil Ahmad, Khan Muhammad, Mi Young Lee, Soonil Kwon, and Sung Wook Baik. Deep features-based speech emotion recognition for smart affective services. *Multimedia Tools and Applications*, 78(5):5571–5589, 2019.

[34] Moataz El Ayadi, Mohamed S Kamel, and Fakhri Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3):572–587, 2011.

[35] Bhagyalaxmi Jena and Sudhansu Sekhar Singh. Analysis of stressed speech on teager energy operator (teo). *International Journal of Pure and Applied Mathematics*, 118(16):667–680, 2018.

[36] Antonio Camarena-Ibarrola, Fernando Luque, and Edgar Chavez. Speaker identification through spectral entropy analysis. In *2017 IEEE international autumn meeting on power, electronics and computing (ROPEC)*, pages 1–6. IEEE, 2017.

[37] Ankita N Chadha, Mukesh A Zaveri, and Jignesh N Sarvaiya. Optimal feature extraction and selection techniques for speech processing: A review. In *2016 International Conference on Communication and Signal Processing (ICCSP)*, pages 1669–1673. IEEE, 2016.

[38] Ali Mirzaei, Vahid Pourahmadi, Mehran Soltani, and Hamid Sheikhzadeh. Deep feature selection using a teacher-student network. *Neurocomputing*, 383:396–408, 2020.

[39] Jie Cai, Jiawei Luo, Shulin Wang, and Sheng Yang. Feature selection in machine learning: A new perspective. *Neurocomputing*, 300:70–79, 2018.

[40] Ismail El Moudden, Mounir Ouzir, and Souad ElBernoussi. Automatic speech analysis in patients with parkinson's disease using feature dimension reduction. In *Proceedings of the 3rd International Conference on Mechatronics and Robotics Engineering*, pages 167–171, 2017.

[41] Zhen-Tao Liu, Qiao Xie, Min Wu, Wei-Hua Cao, Ying Mei, and Jun-Wei Mao. Speech emotion recognition based on an improved brain emotion learning model. *Neurocomputing*, 309:145–156, 2018.

[42] Qipei Mei, Mustafa Gül, and Marcus Boay. Indirect health monitoring of bridges using mel-frequency cepstral coefficients and principal component analysis. *Mechanical Systems and Signal Processing*, 119:523–546, 2019.

[43] Ana Rodríguez-Hoyos, David Rebollo-Monedero, José Estrada-Jiménez, Jordi Forné, and Luis Urquiza-Aguiar. Preserving empirical data utility in k-anonymous microaggregation via linear discriminant analysis. *Engineering Applications of Artificial Intelligence*, 94:103787, 2020.

[44] Chun-Na Li, Yuan-Hai Shao, Wotao Yin, and Ming-Zeng Liu. Robust and sparse linear discriminant analysis via an alternating direction method of multipliers. *IEEE Transactions on Neural Networks and Learning Systems*, 31(3):915–926, 2019.

[45] Xiaowei Zhao, Jun Guo, Feiping Nie, Ling Chen, Zhihui Li, and Huaxiang Zhang. Joint principal component and discriminant analysis for dimensionality reduction. *IEEE Transactions on Neural Networks and Learning Systems*, 31(2):433–444, 2019.

[46] Lawrence R Rabiner. Speech recognition based on pattern recognition approaches. In *Digital Speech Processing*, pages 111–126. Springer, 1992.

[47] Zuzanna Miodonska, Marcin D Bugdol, and Michal Krecichwost. Dynamic time warping in phoneme modeling for fast pronunciation error detection. *Computers in Biology and Medicine*, 69:277–285, 2016.

[48] Usman Khan, Muhammad Sarim, Maaz Bin Ahmad, and Farhan Shafiq. Feature extraction and modeling techniques in speech recognition: A review. In *2019 4th International Conference on Information Systems Engineering (ICISE)*, pages 63–67. IEEE, 2019.

[49] S Shaikh Naziya and RR Deshmukh. Speech recognition systema review. *IOSR J. Comput. Eng*, 18(4):3–8, 2016.

[50] Trishna Barman and Nabamita Deb. State of the art review of speech recognition using genetic algorithm. In *2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI)*, pages 2944–2946. IEEE, 2017.

[51] Lawrence Rabiner. Fundamentals of speech recognition. *Fundamentals of speech recognition*, 1993.

[52] Kennedy Okokpujie, Etinosa Noma-Osaghae, Samuel John, and Prince C Jumbo. Automatic home appliance switching using speech recognition software and embedded system. In *2017 international conference on computing networking and informatics (ICCNI)*, pages 1–4. IEEE, 2017.

[53] Matt Shannon. Optimizing expected word error rate via sampling for speech recognition. *arXiv preprint arXiv:1706.02776*, 2017.

[54] Dat Tat Tran. *Fuzzy approaches to speech and speaker recognition*. PhD thesis, university of Canberra, 2000.