Development of Hausa Acoustic Model for Speech Recognition

Umar Adam Ibrahim[1], Moussa Mahamat Boukar[2]
Computer Science, Nile University of Nigeria
Abuja, Nigeria

Muhammad Aliyu Suleiman[3]
Software Engineering, Nile University of Nigeria
Abuja, Nigeria

***Abstract*—Acoustic modeling is essential for enhancing the accuracy of voice recognition software. To build an automatic speech system and application for any language, building an acoustic model is essential. In this regard, this research is concerned with the development of the Hausa acoustic model for automatic speech recognition. The goal of this work is to design and develop an acoustic model for the Hausa language. This is done by creating a word-level phonemes dataset from the Hausa speech corpus database. Then implement a deep learning algorithm for acoustic modeling. The model was built using Convolutional Neural Network that achieved 83% accuracy. The developed model can be used as a foundation for the development and testing of the Hausa speech recognition system.**

***Keywords*—Acoustic model; Hausa Phonemes; word level; CNN**

I. Introduction

After several years of development and research, the accuracy of automatic speech recognition remains an important research issue. Many things influence the accuracy of a voice recognition application. Speaker and context variations are the most well-known. The use of acoustic modeling helps to improve accuracy. Any speech recognition system relies heavily on acoustic modeling.

The method of building statistic from voice waveform vectors is referred to as acoustic modeling for speech recognition. One of the most prevalent forms of acoustic model is the Hidden Markov Model [1, 2, 3]. Segmental models [4, 5, 6, 7, 8], super-segmental models such as hidden dynamic models [9], neural networks [10, 11], maximum entropy models [12], and hidden conditional random fields [13] are among the other auditory models.

Pronunciation modeling discusses fundamental units of speech sequences. These units include phonetics features. These phonetics features are utilized to present bigger speech units like phrases or words. To achieve noise robustness in speech recognition, acoustic modeling may also include the use of feedback [13].

The acoustic model describes the statistical features of sound occurrences in speech recognition. According to the acoustic model, the likelihood score p(X|W) is calculated. Let's assume the acoustic model component representing an ith word Wi is Yi, the p(X|W) = p(X|Yi) in an isolated-word speech recognition program with N-word vocabulary.

Hidden Markov Model (HMM), Support Vector Machine (SVM), Deep Neural Network (DNN), Artificial Neural Network (ANN), and Convolutional Neural Networks (CNNs) are some of the modeling techniques used [14].

The current work on acoustic model focused on international languages such English, French, Germany etc. Thus acoustic model developed for under-resource language are for Punjabi an India language, Amharic Ethiopian language and others [14].

The researcher implements a supervised Convolutional Neural Network. The Hausa acoustic model dataset was extracted from the Hausa Speech Corpus Database [15]. The researcher outlined the Hausa language alphabet and phonologies with examples.

The rest of the paper is organized as follows: Section II talked about Hausa phonology. Section III provides a review of acoustic modeling. Section IV presented the research methodology. Section VI described the implementation process. In Section V the result obtained was discussed and presented. Lastly, Section VII concludes this paper and mentions potential future work.

The major contribution of this paper is creating word-level phonemes, generating labels for the created phonemes, developing a Hausa acoustic model, testing and validating the developed model.

II. Literature Review

The Hausa language is a Chadic language spoken in West Africa. Hausa has two writing scripts. First, the Ajami writing script is written with the Arabic alphabet, and the Boko scriptwriting is written with the Latin alphabets.

This research is concerned with the Hausa Latin writing script. The Boko script has 36 Latin alphabet, which contain 31 consonants and 5 vowels. The language also has 23 phonetic sounds [16] as shown in Table I.

According to [17] Wurin/Gurbin furunci means a place of articulation. The consonant location in the vocal tract when a blockage occurs between active and passive articulators is known as the place of articulation. There are seven points of articulation for Hausa consonants which are:

Baleɓe meaning Bilabial: this is when the lower lip brushes up against or touches the upper lip.

Bahanƙe meaning Alveolar: when the tip of the tongue meets or touches the alveolar ridge.

Nadɗe-harshe means Retroflex: when the tip of the tongue makes contact with the back of the alveolar ridge.

Ɗan Bayan Hanka meaning Post-alveolar: when the blade of the tongue comes close to or contacts the back of the alveolar ridge.

Bagandɗe meaning Palatal

Bahandɗe meaning Velar: is when the back of the tongue rubs on the velum or soft palate.

TABLE I. HAUSA ALPHABETS

A	B	Ɓ	C	D	Ɗ
E	F	Fy	G	Gw	Gy
H	I	J	K	Ƙ	Kw
Ky	Ƙw	Ƙy	L	M	N
O	R	S	Sh	T	Ts
U	W	Y	`Y	Z	‘
Consonants					
B	Ɓ	C	D	Ɗ	F
Fy	G	Gw	Gy	H	J
K	Ƙ	Kw	Ky	Ƙw	Ƙy
L	M	N	R	S	Sh
T	Ts				
Vowels					
A	E	I	O	U	

TABLE II. HAUSA PHONOLOGY

s/n	Phonology	Phonetic	Sample Word
1	Baleɓe=> Bilabial	/b/	“baka”
		/ɓ/	“ɓarawo”
		/m/	“malam”
		/ɸ/	“fata”
2	Bahanƙe => Alveolar	/d/	“dankali”
		/n/	“talata”
		/r/	“nama”
		/z/	“bara”
		/s/	“zakka”
		/l/	“lada”
		/ts/	“tsawa”
3	Nadɗe-harshe => Retroflex	/ɽ/	“ruwa”
		/dʲ/	“ɗaki”
4	Ɗan Bayan Hanka =>Post-Alveolar	/ʃ/	“shara”
		/dʒ/	“jarida”
		/ʧ/	“canji”
5	Bagandɗe =>Palatal	/j/	“yara”
6	Bahandɗe =>Velar	/g/	“gado”
		/k/	“kaza”
		/ƙ/	“ƙasa”
		/w/	“wasa”
7	Hamza => Glottal	/h/	“hannu”,
		/’/	“sa’a”

Hamza means Glottal: this refers to the closure or constriction of the glottis. Table II shows Hausa phonology, phonetic and sample word according to [17].

Acoustic Modeling (AM) is the first and crucial step in developing a speech recognition platform. The acoustic model generates a link between linguistic and acoustic units. Most of the computations performed in acoustic modeling are because statistical representation and feature extraction affect speech recognition development.

The extracted features are distributed based on a particular sound. This acoustic modeling is done to build a link between the structure of the linguistic model unit and the extracted features.

Different feature extraction methods, like voice production mechanism, and human perception were reported in [18, 19].

The selection of classification algorithms is also a crucial phase in the development of an acoustic model. Many studies on acoustic modeling using various classification algorithms have been published [20]. HMM, ANNs, DNNs, and Sequence to sequence acoustic modeling are some of the classification approaches used by researchers.

AM is associated with a variety of concepts. It necessitates knowledge of acoustic phonetics, microphone, and environment variability issues, gender variations, and dialectal variances. Furthermore, thorough training is required to determine the link between language units and auditory observation [21]. AM is also linked to pronunciation modeling, as well as speaker, environment, and context variability and modeling [22].

III. METHODOLOGY

A generalized acoustic modeling system includes raw data collection, preprocessing, feature extraction, and acoustic modeling. Fig. 1 illustrated the researchers’ acoustic modeling. The pipeline is divided into two components word-level phonology segmentation and word-level phonology labeling. The preprocessing block consists of word-level segmentation and word-level phonology labeling. Furthermore, feature extraction was implemented as the second block in the pipeline. The second block comprises phonemes and labels matched together for modeling purpose. The output of the feature extraction block was fed into the training block for acoustic modeling of the phonology utterances. Finally, a word-level Phonemes based acoustic model was implemented for this research. The word-level acoustic dataset was extracted from the Hausa Speech Corpus database [15].

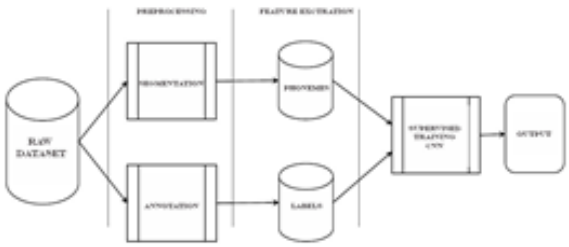

Fig. 1. Hausa Acoustic Model Pipeline.

A. Data Distribution

Five thousand word-level phonemes were extracted from the Hausa Speech Corpus database [15]. Male and female speakers read these words. Table III shows the number of words per literature and the gender of the reader.

TABLE III. DISTRIBUTION OF WORD-LEVEL PHONE AND READER GENDER

S/N	Literature	Words	Gender
1	Gani Gareka	500	Female
2	Iliya dan Maikarfi	700	Female
3	Iki Magayi	700	Male
4	Wani Gari	500	Male
5	Komai Nisa Dare	500	Male
6	Koya Da kanka	700	Female
7	Magana Jarice	700	Male
8	Shehu Umar	700	Male
Total		5000	

The final audio dataset consists of Five thousand phonetically compact word-level phonemes. The phonetically compact words were divided into training, testing, and validation set. The dataset distribution is shown in Table IV.

TABLE IV. SPEECH DATASET

S/N	Division	Percentage	words
1	Training	70%	3500
2	Testing	20%	1000
3	Validation	10%	500

B. Model

Various categorization techniques have been created for audio modeling. HMM, and ANNs are, however, the most extensively utilized algorithms [16]. This work was implemented using deep learning acoustic modeling algorithm. The researchers developed the Hausa acoustic word-level models by implementing a Convolutional neural network (CNN).

IV. IMPLEMENTATION

A. Hardware and Software

The system was built on top of Google Collaborator (Colab) with the following hardware spec:

- CPU: Intel(R) Xeon(R) @ 2.30GHz.
- Disk: Size 79GB, Used-40GB, Available-39GB.
- Memory:Total:~13GB, Free:~10GB, Available:~12GB.

For the software requirement, all necessary modules and dependencies were installed and imported into Colab. Such as OS, Pathlib, MAtplotlib, Numpy, Seaborn, Tensorflow, Keras, and Ipython display. The audio dataset is stored in eight different folders corresponding to each literature named: gani, iliya, jiki, wani, koya, and shehu. The audio clips were extracted and shuffle into a list. The dataset were divided into training, test, and validation set as 70:20:10 ratios, respectively.

B. Preprocessing

The dataset was processed at this phase by creating decoded tensors for the waveforms and labels. Each wav file contains time-series data that is sampled at a specific rate. The amplitude of the audio signal at any given time is represented by each sample. The researchers' WAV acoustic dataset files had amplitude values ranging from -32,768 to 32,767. A 16-bit system was employed. This dataset has a sample rate of 16 kHz.

The shape of the tensor returned by "ft.audio.decode_wav" is the [sample, channels]. The Hausa acoustic dataset contains mono recordings.

Three functions are defined: the first transform the dataset's WAV audio files into tensors. The second is the method that generates labels for each file based on its parent directories. Finally, there is the "get waveform and label" auxiliary function, which ties everything together.

The audio filename is the input and the output is a tuple with audio and label tensors, ready for supervised learning. The audio waveforms were plotted as shown in Fig. 2.

C. Waveforms to Spectrogram

The time domain is used to depict the waveforms in the collection. The waveforms are then converted to spectrograms. The converted spectrogram reveals the frequency variations over time which can be depicted as 2D images. Further, the Short-Time Fourier Transform (STFT) is used to convert from time-domain signals to time-frequency-domain signals. The spectrogram images are then fed into the neural network, which is used to train the model.

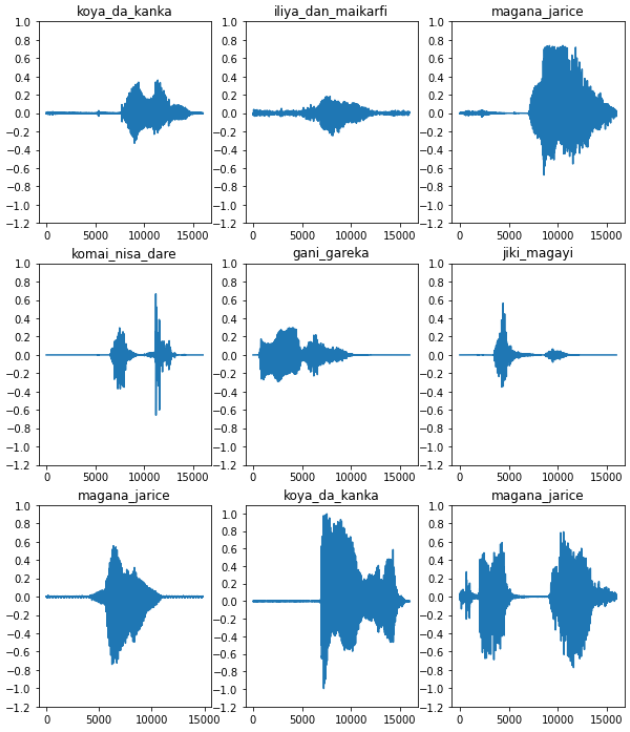

Fig. 2. Audio Waveforms.

Furthermore, a function for converting waveforms to spectrograms was created. The waveforms must be the same length for the spectrograms to have equal dimensions when converted. This can be accomplished y simple zero-padding audio snippets that are less than one second in length using (ft.zeros). In addition, a frame length and frame step options for tf.signal.stft are selected so that the resulting spectrogram "image" is almost square. The STFT generates a complex number of the array that represents magnitude and phase. In this situation, the researcher utilized the magnitude, which can be achieved by running ft.abs on the tf.signal.stft result.

Next, the data was explored by printing the shape of one sample tensorized waveform and the corresponding spectrogram. A function was defined to display a spectrogram. Lastly, sample audio was plotted displaying waveform over time and the corresponding spectrogram (frequencies over time) as shown in Fig. 3.

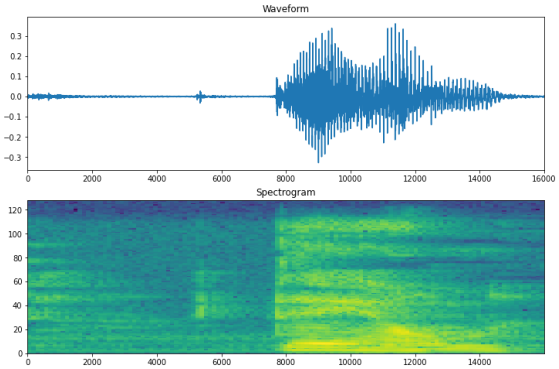

Fig. 3. Waveform and Spectrogram.

A function was then defined to transform the waveform dataset into spectrograms and their corresponding labels as integers. Map function was implemented across the dataset elements. Finally, the spectrograms for different samples of the dataset were examined as shown in Fig. 4.

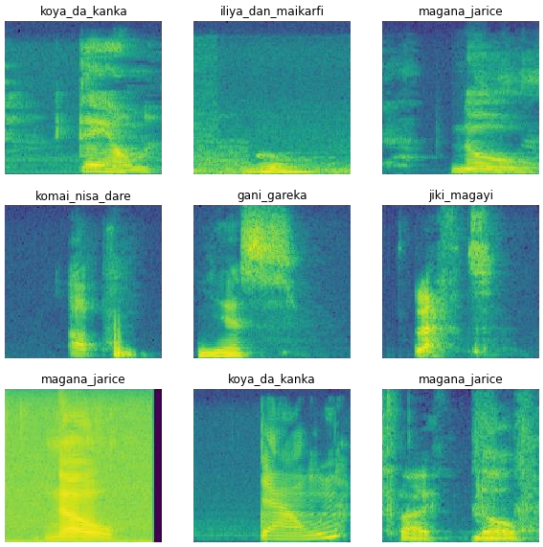

Fig. 4. Samples of Spectrogram.

D. Training the Model

Preprocessing was done on the dataset. Batch training and validation sets were implemented for the model training. Dataset.cache and Dataset.prefetch operations were performed to reduce read latency while training the model.

Since the audio files have transformed into spectrogram images. For modeling a simple Convolution Neural Network was implemented. Keras preprocessing layers were also implemented such as resizing layers to down sample the input to enable the model to train faster. A normalization layer was implemented to normalize each pixel in the image based on its means and standard deviation. The Keras model was configured with Adam optimizer and the cross-entropy loss. The model was trained on 10 epochs for demonstration purposes. The model summary is shown in Fig. 5.

```
Input shape: (124, 129, 1)
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 resizing (Resizing)         (None, 32, 32, 1)         0

 normalization (Normalizatio  (None, 32, 32, 1)        3
 n)

 conv2d (Conv2D)             (None, 30, 30, 32)        320

 conv2d_1 (Conv2D)           (None, 28, 28, 64)        18496

 max_pooling2d (MaxPooling2D  (None, 14, 14, 64)       0
 )

 dropout (Dropout)           (None, 14, 14, 64)        0

 flatten (Flatten)           (None, 12544)             0

 dense (Dense)               (None, 128)               1605760

 dropout_1 (Dropout)         (None, 128)               0

 dense_1 (Dense)             (None, 8)                 1032

=================================================================
Total params: 1,625,611
Trainable params: 1,625,608
Non-trainable params: 3
_________________________________________________________________
```

Fig. 5. Model Summary.

For better analysis, the data was run on several epochs. However, 50 epochs gave better accuracy. The output of the model is shown in Fig. 7.

```
Epoch 1/50
55/55 [==============================] - 11s 195ms/step - loss: 0.2411 - accuracy: 0.9209 - val_loss: 0.5089 - val_accuracy: 0.8260
Epoch 2/50
55/55 [==============================] - 11s 194ms/step - loss: 0.2169 - accuracy: 0.9231 - val_loss: 0.5897 - val_accuracy: 0.8200
Epoch 3/50
55/55 [==============================] - 11s 195ms/step - loss: 0.2296 - accuracy: 0.9180 - val_loss: 0.5663 - val_accuracy: 0.8220
Epoch 4/50
55/55 [==============================] - 11s 195ms/step - loss: 0.1959 - accuracy: 0.9337 - val_loss: 0.6023 - val_accuracy: 0.8300
Epoch 5/50
55/55 [==============================] - 11s 196ms/step - loss: 0.1933 - accuracy: 0.9331 - val_loss: 0.6209 - val_accuracy: 0.8260
Epoch 5: early stopping
```

Fig. 6. Model Output.

V. Result

After running the model with 50 epochs, the summary and output of the model as shown in Fig. 5 and 6. Finally, the training and validation loss curves were plotted to see how the model progressed over time as shown in Fig. 7.

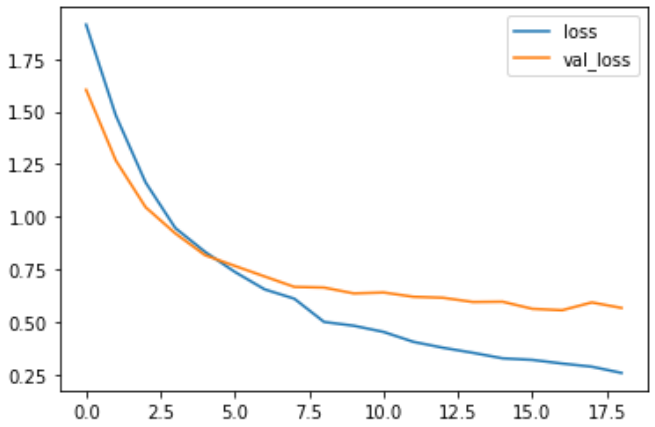

Fig. 7. Train-validation Loss Curves.

To evaluate the model's performance, the model was run on the test set. The accuracy of the test was 83%.

To see how successfully the model classified each auditory word in the test set, a confusion matrix was plotted as displayed in Fig. 8.

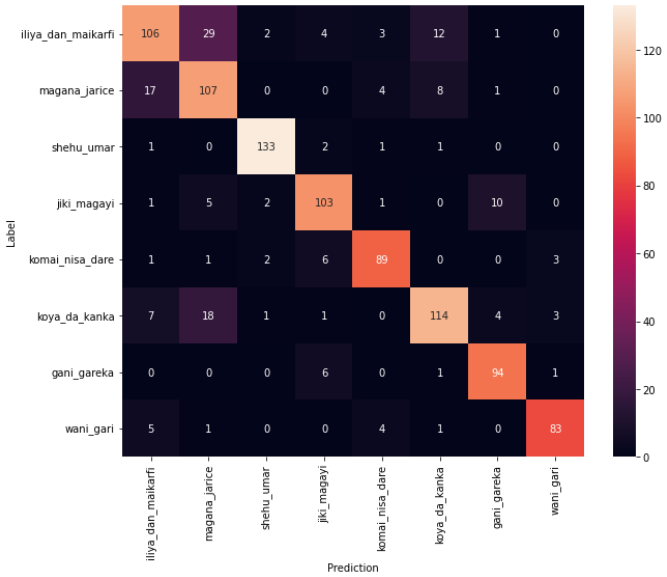

Fig. 8. Confusion Matrix.

VI. Conclusion

This article started with an introduction to the acoustic model. Stated the previous and current techniques used for the development of an acoustic model for speech recognition and then introduces Hausa phonology, its examples, and alphabets. The researchers created word-level phonemes from the Hausa Speech Corpus database.

With the rise of deep learning algorithms for acoustic model development, the researchers implemented CNN base acoustic model development for the Hausa language. Hausa is a low-resourced and under-resource language. The goal is to create an acoustic model for the Hausa language.

The created model can be used for the development of Hausa speech recognition system. The outputs suggest that the model recognizes Hausa phonetic with 83% accuracy. The researchers' future work is to develop a language model for the Hausa language. The acoustic model and language model would be linked to developing an automatic speech recognition system for the Hausa language.

References

[1] L. Baum, "An inequality and associated maximization technique occurring in statistical estimation for probabilistis functions of Markov process," Inequality III, 1972, pp. 1–8.

[2] J. Baker, "Stochastic modelling for automatic speech recognition in D.R Reddy," In D.R Redddy, (ed), Speech Recognition, Academic Press, New Yok, 1975.

[3] F. Jelinek, "A fast sequential decoding algorithm using a stack," in IBM journal of Research and Development, vol. 13, pp. 675–685, 1969.

[4] A. Poritz, "Hidden Markov Models: A guided tour," in Proceedings of the Internatonal conference on acoustic, Speech and Signal processing, Vol. 1, pp. 1-4 1998.

[5] L. Deng, "A stochastic model of speech incorporating hierarchical nostationary," IEEE Transaction on speech and audio processing, Vol. 1(4), pp.417-475, 1993.

[6] L. Deng, M. Aksamanovic, X. Sun, and C. Wu, "Speech recognition using hidden markov models with polynimial regression function as nostationary states," IEEE Transaction on speech and audio processing, vol. 2, pp. 507–520, 2004.

[7] M. Ostendorf, V. Digalaskis, and J. Rohlicek, "From HMMs to segment models: A unified view of stochastic modelling for speech recognition," IEEE Transaction on speech and audio processing, vol. 4, pp. 360–378, 1996.

[8] J. Glass, "A probabilistic framework for segment-based speech recognition," in M. Russell and J. Bilmes(eds), New computational paradigms for acoustic modelling in speech recognition computer, speech and language (special issue), 17(2-3), pp. 137–155, 2003.

[9] L. Deng, D. Yu, and A. Acero, "Structured speech modelling," IEEE Transaction on speech and audio processing, (special issue on rich transcriptin), vol. 14(5), pp. 1492–1504, 2006.

[10] R. Lippma, "An introduction to computing with neural nets," IEEE ASSP Magazine, 4(2), pp. 4–22, 1987.

[11] N. Morgan, et. al, "Pushing the envelop-aside," IEEE signal processing magazine, pp. 81–88, 2005.

[12] Y. Goa, and J. Kuo, "Maximum entropy direct models for speech recognition," IEEE Transaction on speech and audio processing, vol. 14(3), pp. 873–881, 2006.

[13] A. Gunawardana, and W. Bryne, "Discriminative speaker adaptation with conditional maximum likelihood linear regression," in proceedings of the EUROSPEECH, aalborg, Denmark, 2001.

[14] S. Bhatt, A. Jain, and A. Dev, "Acostic modelling in speech recognition: A system review," Internation Journal of Advance computer science and applications, vol. 11, No. 4, 2020.

[15] U.A. Ibrahim, M.M. Boukar, and M.A Suleiman, "Development of Hausa dataset a baseline for speech recognition," Dat in brief, 40, 2022.

[16] https://wisc.pb.unizin.org, "Hausa Alphabet" [online]. Available:https://wisc.pb.unizin.org/lctlresources/chapter/hausa-alphabet [Accessed: 20-Nov-2021].

[17] www.amsoshi.com, "Phonology 1(Wurin/Gurbin Furuci(Place of Articulation))" [online]. Available: https://www.amsoshi.com/2020/02/alh-203-hausa-phonology-1-wuringurbin.html. [Accessed: 20-Nov-2021].

[18] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," Journal of the Royal Statistical Society, vol 39(1), pp. 1-21, 1977.

[19] H. Hermansky, "Perceptual linear predictive (PLP) Analysis of speech," J. Acoust. Soc. Am., vol. 87, no. 4, pp. 1738-1752, 1990, doi:10.1121/1.39923.

[20] S. Bhatt, A. Dev, and A. Jain, "Confusion analysis in phoneme based speech recognition in Hindi," J. Ambient Intell. Humaniz. Comput., no. 0123456789, 202, doi: 10.1007/s12652-020-01703-x.

[21] M. J. F. Gales, S. Watanabe, and E. Fosler-Lussier, "Structured Discriminative models for speech recognition: An overview," IEEE Signal Process. Mag., vol. 29, np. 6, ppp. 70-81, Nov 2012. Doi: 10.1109/MSP.2012.2207140.

[22] R. K. Aggarwal and M. Dave, "Acoustic modeling problem for automatic speech recognition system: Advances and refinements(part II)," Int. J. Speech Technol., vol. 14, no. 4, pp. 309-320, 2011, doi:10.1007/s10772-011-9106-4