Comparison of Khasi Speech Representations with Different Spectral Features and Hidden Markov States
Bronson Syiem, Sushanta Kabir Dutta, Juwesh Binong, and Lairnalkpam Joyprakash Singh


Articles you may be interested in


Follow JEST WeChat public account for more information
**Comparison of Khasi Speech Representations with Different Spectral Features and Hidden Markov States**

Bronson Syiem* | Sushanta Kabir Dutta | Juwesh Binong | Lairenlakpam Joyparaksh Singh

**Abstract**—In this paper, we present a comparison of Khasi speech representations with four different spectral features and novel extension towards the development of Khasi speech corpora. These four features include linear predictive coding (LPC), linear prediction cepstrum coefficient (LPCC), perceptual linear prediction (PLP), and Mel frequency cepstral coefficient (MFCC). The 10-hour speech data were used for training and 3-hour data for testing. For each spectral feature, different hidden Markov model (HMM) based recognizers with variations in HMM states and different Gaussian mixture models (GMMs) were built. The performance was evaluated by using the word error rate (WER). The experimental results show that MFCC provides a better representation for Khasi speech compared with the other three spectral features.

**Index Terms**—Acoustic model (AM), Gaussian mixture model (GMM), hidden Markov model (HMM), language model (LM), linear predictive coding (LPC), linear prediction cepstral coefficient (LPCC), Mel frequency cepstral coefficient (MFCC), perceptual linear prediction (PLP).

1. **Introduction**

Speech is the easiest mode of communications between individuals and in fact it is the most efficient form of exchanging information among humans. The perception of speech looks simple for humans but in reality it is quite complicated. The same word of a language can be pronounced differently by different people depending upon the region they have been living in. Also, every individual has unique pitch and formant. Hence, the representation of the speech signal is required in order to understand different features associated with and to preserve the information needed to determine the phonetic identity of speech. Generally, the representation of speech is specified by a model description including the parameters and coefficients of the...
signal. Based on these parameters, speech representations can be classified into two groups, namely the parametric representation (based on model parameters, e.g., pulse code modulation) and the non-parametric representation (only consisting of signal coefficients, e.g., sinusoids). The representation of speech in a complete and compact form is of great importance in an automatic speech recognition (ASR) system. The usage of an appropriate representation of the speech signal yields a more efficient speech coding system, which could improve the quality of speech synthesis and the performance of recognizers. Speech recognition systems generally assume that the speech signal is a realization of some messages encoded as a sequence of one or more symbols. To realize the reverse operation of recognizing the underlying symbol sequence from the spoken utterance, the continuous speech waveform is first converted to a sequence of equally spaced discrete parameter vectors. This sequence of parameter vectors is assumed to form an exact representation of the speech waveform for the duration covered by a single vector (typically 10 ms or so). Much research has been carried out in this area to represent speech, however, majority of the work done uses only a few different signal representations. The cepstrum coefficient (CC) is a representation which is the most commonly used in speech recognition systems.

Using the canonical representation of speech, the separation of voiced and unvoiced sounds with the TIMIT database was performed in the past. The experiment focused on this separation with a high-efficient result was achieved. However, it did not perform the investigation of speech recognition. Other related studies using the spectral representation with combined pitch-synchronous acoustic features were also done, in which the result showed a relatively low word error rate (WER). However, the experiment was only performed with few spectral features. The speech representation based on phase locked loop (PLL) features of voiced sounds with the TIMIT database showed a higher recognition rate when it was measured over a range of high noise, yet issues remain to cover the unvoiced segments. The phone recognition using three spectral features (linear prediction cepstral coefficient (LPCC), perceptual linear prediction (PLP), and Mel frequency cepstral coefficient (MFCC)) on the Manipuri language was performed with three different modes of conversation. The results showed a slight improvement. However, they did not perform the analysis with varying hidden Markov model (HMM) states and mixture models. In this work, the speech recognition would be performed with varying HMM states and mixture models, and meanwhile more spectral features will be used.

Despite major advancements have been achieved in the state-of-the-art speech recognition systems, there still exists a huge and challenging task, particularly for those with low resource languages. Till date, very limited research has been conducted with regard to to Khasi speech recognition and as far as our knowledge is concerned, no Khasi speech recognition with different HMM states has been performed. In this paper, we investigate the Khasi speech representation using different spectral features with three different HMM states, which is an extended version of our previous work.

2. Khasi Language

The Khasi language is one of the Austro-Asiatic languages under the Monkhmer branch and is spoken by the natives of the state Meghalaya. Depending on the geographical location and local residents, the dialects vary to some extent. Based on this, Bareh proposed a total of 11 Khasi dialects. According to the research of Nagaraja et al., Khasi dialects are comprised of four major dialects including Khasi Proper (from Sohra), Pnar, Lyngngam, and Wa. Amongst these four dialects, Khasi Proper is the standard Khasi dialect and hence it is chosen to be studied in this work.
3. HMM State

HMM can be considered as a stochastic, finite state machine with an unobserved state used for modeling speech utterance\(^{[12]}\). In HMM, the state is hidden, however, the output that depends on the state is assumed to be visible. Since the state is hidden, the only parameters for HMM are the transition probabilities \(a_{jk}\) and the emission probabilities (also called output probabilities)\(^{[13]}\). Fig. 1 shows the general state diagram of HMM with the \(j\)th state and the \(i\)th output probability distribution function. The states are fed forward, in other words, the states can stay in itself or move from the left-right direction but not in the reverse (right-left) direction. The starting state (\(q_1\)) and the exit state (\(q_j\)) are assumed to be non-emitting states, which means these states do not generate observations. The states (\(q_2\) to \(q_{j-1}\)) are considered to be emitting states with output probabilities of \(b_1\) to \(b_i\). These output probabilities generate the observation \(o_i\), (the acoustic feature vector). Each observation probability is represented by a Gaussian mixture density\(^{[14]}\). In this work, the 3, 5, and 7 HMM states have been considered.

4. Spectral Features

In ASR systems, the feature plays a crucial factor. In this work, four different spectral features including linear prediction coding (LPC), LPCC, PLP, and MFCC have been used and they are briefly explained as below.

4.1. LPC

LPC is the method of estimating basic speech parameters (e.g., pitch and formant) and used for representing speech with a low bit rate\(^{[15]}\). It is based on the fact that a present speech sample can be approximated as a linear combination of past speech samples.

4.2. LPCC

LPCC is the derived feature of LPC, which is one of the most popular techniques used for the speaker recognition system\(^{[16]}\). To compute the LPCC feature, the LPC spectral envelope is first computed. The technique is similar to the MFCC computation. In MFCC, pre-emphasis is applied to the speech waveform, whereas in LPCC it is applied to the spectrum of the input speech signal\(^{[17]}\).

4.3. PLP

This technique is similar to LPC. However, in PLP the spectral characteristics are transformed to match the human auditory system characteristics, hence PLP is more adaptable to human hearing compared with LPC. The main difference between these two techniques is that PLP assumes the all-pole transfer function of the vocal tract with a specified number of resonances\(^{[18]}\).

4.4. MFCC

This feature is defined as the representation of the short-term power spectrum of a speech sound. It is based on a linear cosine transform of a log power spectrum on a non-linear Mel scale of frequency. In MFCC, the frequency bands are equally spaced on the Mel scale. This characteristic feature enables MFCC to
approximate the human auditory system’s response more closely compared with the linearly spaced frequency bands used in the normal cepstrum\textsuperscript{[18]}.

5. Database Description

The collection of speech data can be made from different modes (e.g., read mode, lecture mode, or conversation mode). The speech data used for this experiment are read-based, uttered by 120 Khasi native speakers. The speech data were recorded in a laboratory by using the Zoom H4N Handy portable digital recorder with a sampling rate of 16 kHz, where each speaker was given 50 unique sentences to read. The detailed description of the speech database is shown in Table 1.

6. Experimental Setup

The experiment was carried out in the Ubuntu 14.04 long term support (LTS) platform using the hidden Markov model toolkit (HTK). The features were extracted by applying a Hamming window of 25-ms size with a frame shift of 10 ms. The extracted spectral features were used for creating the acoustic model (AM), whereas the transcription labeled files were used for creating the language model (LM). The numbers of wave files used for training and testing were 6000 and 1800, respectively. Different AMs were built by using 3, 5, and 7 HMM states to represent each phone. Then the global mean and variance per HMM state were calculated by using the predefined prototype along with the acoustic vectors and transcriptions of the training data set\textsuperscript{[19]}. The optimal values for the HMM parameters (transition probability, mean and variance vectors for each observation function) were re-estimated once an initial set of models was created. The pronunciation dictionary/lexicon was built by splitting the word sequence into the phone sequence. Table 2 shows the illustration of the dictionary used in the experiment. Each phone acted as an acoustic unit that was further used for training the HMM model\textsuperscript{[20]}. LM was used for creating $n$-gram corresponding to the text transcription. In this experiment, we used bi-gram LM. Finally, the decoder captured the distinct sounds of each word and produced the output. It was further matched with the trained HMM for each sound then the phone sequence was determined like in a pronunciation dictionary. Finally, the word was recognized. The detailed process is schematically shown in Fig. 2.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sampling rate</td>
<td>16 kHz</td>
</tr>
<tr>
<td>Number of sentences</td>
<td>6000</td>
</tr>
<tr>
<td>Number of male speakers</td>
<td>65</td>
</tr>
<tr>
<td>Number of female speakers</td>
<td>55</td>
</tr>
<tr>
<td>Duration of wave file</td>
<td>5 s/wave file to 7 s/wave file</td>
</tr>
<tr>
<td>Age of speakers</td>
<td>20 years to 50 years</td>
</tr>
<tr>
<td>Total number of speakers</td>
<td>120</td>
</tr>
<tr>
<td>Total duration of speech data</td>
<td>10 h</td>
</tr>
</tbody>
</table>

Table 2: Illustration of lexicon used in the experiment

<table>
<thead>
<tr>
<th>Word sequence</th>
<th>Phone sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>ngan wan</td>
<td>nga n w a n</td>
</tr>
<tr>
<td>nga la kular</td>
<td>nga l a k u l a r</td>
</tr>
</tbody>
</table>

Fig. 2. Schematic representation of speech recognition.
7. Results and Discussion

In this study, we have used four different spectral features with three different coefficient dimensions. HMM-based recognizers were built with different HMM states and different Gaussian mixture models (GMMs). From the experiment, we investigated the following three cases.

7.1. Case 1

In this case, the 13-dimensional spectral features were used to model HMMs with three different HMM states (3, 5, and 7) and GMMs. The results are shown in Fig. 3. Using the LPC feature, we observed a relatively low reduction in WER irrespective of the variations in HMM states and GMMs. With the LPCC feature, more reductions of WER were noticed with respect to the changes in HMM states and GMMs. Furthermore, using the PLP and MFCC features, much more reductions of WER were presented as compared with the formal two features. For each feature, increasing the mixture components, more improvements in recognition were observed particularly for those with LPCC, PLP, and MFCC. However, a further increase in the HMM states to 7 shows a reduction in recognition performance.

7.2. Case 2

In this case, we used 26-dimensional features to build the models. By increasing the feature dimension, an improvement in recognition accuracy was observed. However, no improvement was obtained by using the LPC feature when increasing the feature dimension as well as HMM states and mixture components. Like in the first case (Case 1), there was a reduction in recognition performance when further increasing HMM states to 7, but with the increase of mixture components, it yielded good performance. The detailed change of WER can be observed in Fig. 4.

7.3. Case 3

In the last case, we further increased the feature dimension to 39. The results can be observed from Fig. 5. With further increasing the feature dimension, a more obvious improvement was raised in recognition than those in Case 1 and Case 2. However, by using the LPC feature, it showed poor performance. Besides 5 states, the variations of HMM states also did not show much enhancement in recognition.

In all the three cases, it was observed that the LPC and LPCC features resulted in poor performance. This may be due to the high noise sensitivity and the lack of using the human auditory feature during the feature extraction process as stated in [19]. Further, it was observed that increasing the feature dimension could
improve the performance. This possibly results from the insufficient coefficient to capture the phones under a low feature dimension. In order to make HMM more accurate, the mixture size must be increased\(^7\) and, from the experiment, the optimal mixture size was found to be 64. Furthermore, though the increase of HMM states led to the increase of acoustic likelihood, it was found that when exceeding 5 states, no improvement was produced. This may be due to the short sequence of phones considered for creating the HMM model. The best performance was achieved by using 39-dimensional MFCC features with 5 HMM states, with the recognition accuracy of 90.36%.

8. Conclusion

In this experiment, we performed a comparison of Khasi speech representations with different spectral features and HMM states. The performance evaluation was done by building different HMM-based recognizers using the HTK ASR toolkit for different spectral features with different feature dimensions, HMM states, and GMMs. The experimental results showed the MFCC feature provided a better representation of Khasi speech. This study reveals that the ASR system does not only depend on model parameters but also the type of features used. One major outcome which can be derived from this work is that it has provided baseline information of the different features and HMM states. For future work, since ASR is a data driven system, more speech data can be incorporated and also the neural network classifier can be applied to improve the system performance.

Disclosures

The authors declare no conflicts of interest.

References


Bronson Syiem received his M.Tech. degree in electronics and communications engineering from North-Eastern Hill University, Shillong in 2015. He is currently pursuing his Ph.D. degree with the Department of Electronics and Communication Engineering, North-Eastern Hill University. His areas of research include speech processing & its applications and pattern recognition.
Sushanta Kabir Dutta received the B.E. degree from Jorhat Engineering College, Jorhat in 1999. He received the M.Tech. degree from Manipal Institute of Technology, Manipal in 2006 and the Ph.D. degree from North-Eastern Hill University in 2018. Now, he is an associate professor with the Department of Electronics and Communication Engineering, North-Eastern Hill University. His research interest is focused on the issues of speech signal processing, pertaining to language identification, image processing, and machine learning.

Juwesh Binong received his B.E. degree in electronics and communications engineering from Bhavanagar University, Bhavanagar in 1999, the M.S. degree in information technology from Sikkim Manipal University, Gangtok in 2007, and the Ph.D. degree in computer science and engineering from Tezpur University, Tezpur in 2017. Now, he is an assistant professor with the Department of Electronics and Communication Engineering, North-Eastern Hill University. His research interests are in the areas of artificial intelligence and its applications.

Lairenlakpam Joyprakash Singh received his M.Tech. degree from Tezpur University, Tezpur in 2000 and the Ph.D. degree in engineering from Jadavpur University, Kolkata in 2006. Now, he is a professor with the Department of Electronics and Communication Engineering, North-Eastern Hill University. His research interests are in the general areas of signal processing and its applications.