Application of Triphone Clustering in Acoustic Modeling for Continuous Speech Recognition in Bengali

Pratyush Banerjee, Gaurav Garg, Pabitra Mitra, Anupam Basu
Communication Empowerment Lab, IIT Kharagpur, India
{pratyushb, gaurav.garg28, pabitra, anupambas}@gmail.com

Abstract

The performance of the acoustic models is highly reflective on the overall performance of any continuous speech recognition system. Hence generation of an accurate and robust acoustic model holds the key to satisfactory recognition performance. As phones are found to vary according to the position of occurrence within a particular word, context information is of prime importance in acoustic modeling of phonetic signals. In this paper we look at the effect of triphone-based acoustic modeling over monophone based acoustic models in the context of continuous speech recognition in Bengali. Keeping in mind the lack of training resources for triphone-based acoustic modeling in Bengali, we have also described herein, the method of generating triphone clusters using decision tree based techniques. These triphone clusters have then been used to generate tied-state triphone based acoustic models to be used in a continuous speech recognizer.

1. Introduction

Continuous speech recognition has been an area of active research for quite some time now. However when compared to languages like English or French, the state of speech research involving Indian languages is yet to gain momentum. Although some amount of effective research has gone into the development of speech recognizers in Hindi [1] and some south Indian Languages [2], the research scenario for Bengali language is far from satisfactory. In course of our effort to create a continuous speech recognizer for Bengali, we came across certain issues involving the generation of robust acoustic models for Bengali.

Speech recognition involves the conversion of a spoken signal into its corresponding most likely written form. The main task is achieved by comparing certain features of the input signal with respect to some pre-trained acoustic models [3]. The modeling units chosen for acoustic modeling in our context were phones. We designed a phone set comprising of 48 phones capturing nearly all the sounds available in standard colloquial Bengali utterance to start with.

In the following section we will briefly discuss about the standard acoustic modeling techniques that we have used to generate models for Bengali highlighting the reasons for the choice of the approaches. Section three mentions the steps involved in generating triphone clusters. Finally in section 4 we provide the experimental results from our work followed by future course of action and conclusion.

2. Acoustic Model in Continuous speech recognition

The acoustic model in a speech recognition engine produces the basic units of speech in the written form with respect to a particular input signal. An input signal is spliced up into overlapping timeframes of 10ms with a 5 ms overlap. Then from these individual frames, 39 MFCC [3] features are extracted. These set of features are then compared with respect to the trained acoustic model. For our purpose we started by creating of single Gaussian monophone Hidden Markov Models (HMM) [4] for every phone in our phone set.

In course of our efforts in building a speech recognition engine we have used the tools provided in the Hidden Markov Model Toolkit or HTK [5]. The HTK provides a set of APIs which are specially suited for the purpose of using HMMs to design configure and evaluate a speech recognizer [5].

2.1 Monophone based acoustic models

The process of creation of monophone acoustic models starts with preparation of the training and testing data. This data comprises of utterance recordings by multiple speakers and the corresponding transcripts encoded using the chosen phone set for the language. These transcripts along with the recordings are fed to the training module which utilizes Baum-Welch Re-estimation [3] in order to create HMMs of all the phones occurring in the training data. The process starts with a default prototype HMM for every phone which is iteratively tuned according to the input data and transcriptions. Creation of the monophone HMMs however require specifying the number of states prior to training. Our experimentations suggested utilizing 5-state HMMs for the purpose of acoustic modeling.

However the monophone based models cannot capture the variation of a phone with respect to the context. Phones are found to vary depending on the preceding and succeeding phones [6] and this aspect needs to be captured within the acoustic models to improve performance.

2.2 Triphone based acoustic models

Looking at the performance issue involved in using monophone-based acoustic models along with the motivation of incorporation of context information within the phone models, led us to try triphone based models for acoustic modeling. Triphone based modeling involve capturing the context information within the phone models at the cost of training much more basic units [6]. Both left and right context information is used to capture the dependency of the observations to represent continuity and co-articulation effects in continuous speech [5]. In our context the phoneme model is conditioned by the phones just preceding and succeeding the current phone. We have utilized the *within-word-context-expansion* technique [5] wherein individual phones within words are modeled by taking into account their left and right context and the phonemes occurring in the word boundaries are modeled as diphones.

The major steps involved in triphone modeling are the same as that of monophone modeling except that it requires far more training data for successful model generation. Ideally, if K words are involved in the successful training of monophone models, training of triphone models would require K^3 words. This remained a huge problem since such abundant training data is not available in Bengali. Even if sufficient training data is provided there is no guarantee that every possible triphone would be occurring in that training data, or some triphones might occur with such low frequency that proper estimation is not possible for them. All these factors make triphone modeling a difficult proposition for Bengali.

3. Clustering of triphones

In order to work around the above mentioned problems for triphone modeling, triphones are clustered together on the basis of the phonetic similarities [7]. This reduces the effective number of units we would try to model eventually. In other words instead of modeling the every triphone, we modeled the phoneme clusters. The clustering technique used for the purpose is decision tree-based clustering [8]. The following figure illustrates an example of the phonetic tree which we have used to cluster the Bengali phone set.

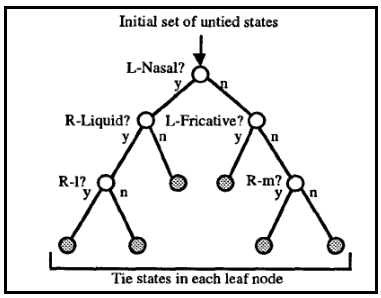

Fig 1: Example of a phonetic tree

We have used phonetic decision trees wherein every node is associated with a particular question regarding the phonetic structure of the phones in context of the current phone. We have clustered all the triphones which have phonetically similar context and created tied state HMM models on them.

3.1 Phonetic Decision Tree Construction

In order to construct the phonetic tree we start with the Bengali phone set comprising of 48 chosen phones and the annotated training data set which constitutes of the utterance recordings and their corresponding phonetic transcriptions. A tree is constructed for each phone in the phone set in order to associate all the states of the associated triphones. The tree topology and the splitting criteria, i.e. the phonetic questions associated with every node are chosen to maximize the likelihood of the training data given the tied states. At the same time it is also ensured that sufficient data associated with each tied state exists, such that effective estimation of the HMM parameters is

possible. Once such tree has been constructed, an unseen triphone can be synthesized simply by identifying the terminal node depending on the context and utilizing the tied states associated with the node to construct the model of the triphone.

Each tree is constructed using a top down sequential optimization procedure [9, 10]. Initially all the states to be clustered are placed at the root of the tree assuming that all the states are tied in that node and the log-likelihood of the training data is computed. This node is then split up into two child nodes on the basis of the question which provides a partition having the maximum log-likelihood increase. This process is repeated at every node, until the increment in log-likelihood falls below a threshold. In order to ensure that every terminal node has sufficient data associated with it, we use a minimal occupation count for each node is also maintained.

It is apparent from the procedure that the most important aspect in constructing the decision trees is the design of the phonetic questions. These questions are primarily based on the phonetic properties of phones used in the context. Hence the process of creating a set of questions required looking into the phonetic table for the Bengali phones and identifying the specific phonetic properties that might be utilized to create the question set. Since phonetic properties are specific to individual languages, hence the already designed questions sets for languages like English or German were of little use in our case. The following section elaborates on our efforts in creating the question set for Bengali phones.

3.2 Creation of the Phonetic Question Set

Creating the set of questions with respect to the Bengali phonetic table and the specific phonetic properties of the Bengali phones involved classifying all the phones on the basis of there phonetic properties.

In Bengali language the phones can be broadly be classified according to the place of excitation, manner of excitation and type of excitation. On the basis of these criteria we identified the articulatory and phonetic properties that formed the basis of designing the phonetic question set. The following properties were eventually used to classify the Bengali phones and generate the phonetic questions.

- Class of phone: Vowel, consonant or semi-vowel
- Length of phone: Short or long.
- Position of jaw: High, medium or low.
- Position of articulation: Front, central or back.
- Voicing: Voiced or unvoiced.
- Continuance: Continuant or non-continuant.
- Rounding of lips: Round or not.
- Tension in cheeks: Tense or lax.
- Manner of articulation: Stop, affricate, fricative, nasal, liquid
- Point of articulation: Bilabial, labial, velar, alveolar, labiodental, alveopalatal or interdental.
- Stridency: Strident, non-strident, or unstrident.
- Zone of articulation: Anterior or non-anterior.
- Position of front of the tongue: Whether the consonant is coronal or not.
- Degree of muscular effort: Fortis, lenis, or neither.
- Aspiration: Whether aspirated or non aspirated

Once the set of properties were identified we went about classifying each phone in the Bengali phone set according to the properties. The classification was done manually and once every property had associated phones with it, the question set was decided on the basis of these properties.

```
QS  "R_Stop"        {*+k,*+kh,*+g,*+gh,*+T,*+th,*+D,*+Dh,
                      *+T,*+th,*+d,*+dh,*+p,*+ph,*+b,*+bh }
QS  "R_Nasal"       {*+~N,*+n,*+m }
QS  "R_Fricative"   {*+s,*+sh,*+h,*Sh }
QS  "R_Affricate"   {*+ch,*+chh,*+j,*jh,*+y }
QS  "R_Lenis"       {*+g,*+gh,*+j,*+jh,*+D,*+Dh,*+d,*+dh,
                      *+b,*+bh }
```

Fig 2: Example of the phonetic question sets

The above figure illustrates a couple of questions we designed for the purpose of phonetic classification. All the questions were of the form, “Does the left or right context of the phone belong to a particular property set P”. Considering an example, the question R-Nasal actually defined “Does the right context of the phone belong to the nasal property set”.

Although the basic nature of the question set designed for Bengali phones is similar to that used for English, the questions were based on the Bengali phonetic properties which rendered them different from the English question set. Using this question set we were able to generate phonetic trees which in turn enabled us to generate tied-state triphone models.

4. Experiments and Results

Using the techniques mentioned so far, we have conducted several experiments with the acoustic model of the Bengali continuous speech recognizer. We used about 4000 recorded utterances from 18 male and 4 female speakers, as the training data for the acoustic models. 600 different sentences from the same speakers, which wasn’t a part of the training data, were used as the testing data. The metric that were used for

the purpose of testing the system were percentage of correct words and accuracy [6]. The testing

Our experiments involved training both monophone based and triphone based models with the same training data and testing the whole recognition system with respect to the same testing data set. The following figure clearly illustrates the results we have obtained in course of our experiments. It is evident from the chart that tied-state triphone based models clearly outperform monophone based models.

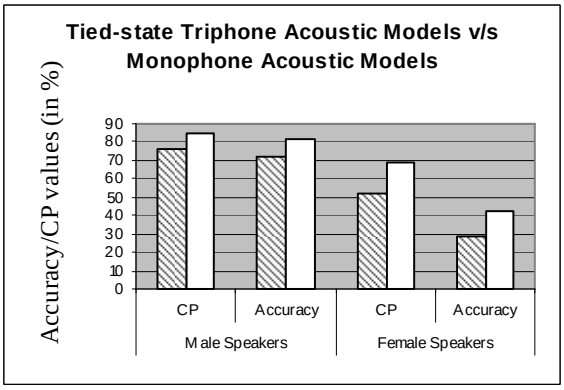

Fig 3: Comparative performance of tied-state triphone models (white bars) and monophone models (shaded bars)

As seen in the figure above, considering separate experiments for male and female speakers, we obtained an average recognition rate (PC) of 76.33% and accuracy of 71.62% for monophone based acoustic models. For female speakers, the lack of sufficient training data caused the figures to drop to 52.34% and 28.67% respectively.

For simple triphone based acoustic modelling without using state tying, recognition rate was found to drop considerably to about 28% for male speakers. The main reason for the poor performance of such models may be attributed to lack of sufficient training data for the purpose.

Tying the triphone states, using a 1900 node 10 level decision tree, resulted in reduction of 18000 separate acoustic states to 1900 tied-states. This experimentation resulted in an average recognition rate of 84.9% and accuracy of 81.16% for male speakers. For female speakers the recognition rate was 69.14% with 42.86% accuracy.

Hence looking at the results we see an average improvement of nearly 13% for male speakers and 40% for female speakers. The fact that tied-state triphone models can synthesize unseen triphones is highlighted in the spectacular improvement in the female recognition figures despite being trained with only 500 sentences. The improvement in the male recognition figures also suggests the suitability of the tied-state triphone based modeling technique for the purpose of speech recognition.

5. Conclusions and Future Work

In this paper we have shared our observations on the experiments we conducted with monophone and triphone based acoustic models for speech recognition in Bengali. Our results clearly indicate the improvement in recognition accuracy by using tied-state triphone models. Hence we can conclude with some rationality that tied-state triphone modeling technique results in far more accurate and better acoustic models even in a poor resource scenario.

Going towards the future we intend to utilize more context information, not only dependent on the phonetic properties of the context but also looking at the positional information of the phones with respect to the words. Using speaker adaptation techniques might also improve the accuracy for the acoustic models in the current poor resource scenario for Bengali.

References

[1] M. Kumar, N. Rajput and A Verma, "A Large Vocabulary Continuous Speech Recognition System for Hindi" *IBM Journal of Research and Development*, vol. 48, number 5/6, 2004.

[2] C.S. Kumar and F. S. Wei, "A Bilingual Speech Recognition system for English and Tamil" *ICICS PCM*, December 2003.

[3] L. R. Rabiner , et. al. "Fundamentals of Speech Recognition", *Prentice Hall Inc.*, 1993

[4] F. Jelinek, "Statistical Methods for Speech Recognition", *Cambridge, MA: MIT Press*, 1997

[5] Steve Young, "The HTK Book", Microsoft Corporation and Cambridge University Engineering Department (CUED), Ver 3.4, Dec, 2006

[6] L.R. Bahl, R. Bakis, P.S. Cohen, A.G. Cole, F. Jelinek, B.L. Lewis, and R.L. Mercer. Further results on the recognition of a continuously read natural corpus. In *ICASSP*, pages 872–875, 1980.

[7] Young SJ (1992). "The General Use of Tying in Phoneme-Based HMM Speech Recognisers." Proc ICASSP, Vol 1, pp569-572, San Francisco

[8] L. Breiman, J.t. Friedman, R.A. Olshen, C.J. Stone, "Classification and Regression Tree", Wadsworth Statistlcs Probability Series, Behnont, CA, 1984.

[9] Kannan A, Ostendorf M, Itohlicek JR (1994). "Maximum Likelihood Clustering of Gaussians for Speech Recognition" to appear, IEEE 'lh'ans on Speech and Audio Processing.

[10] Odell JJ. (1992) "The Use of Decision Trees with Context Sensitive Phoneme Modelling" MPhil Thesis, Cambridge University Engineering Department