Transcription variants effects on the continuous Punjabi language ASR system

Jyoti Guglani, A.N. Mishra

***Abstract*—**The performance of Continuous Punjabi language with different transcription variants is analyzed. The Performance of Automatic Speech Recognition system is much affected by the variations in terms of word error rate. The effect of training of acoustic model through canonical lexicon and lexicon with variants on the speech recognition system evaluated. The performance of the system with both training sets was investigated, and results were compared. Both monophone and triphone models of Speech Recognition were used to analyze the performance of the system. The language used for Automatic Speech Recognition System is the Punjabi language. The Continuous Punjabi language is chosen for the recognition system.

***Keywords*—** Automatic speech recognition system, Canonical lexicon, Continuous Punjabi language, Word error rate.

I. Introduction

The main challenges with Automatic Speech Recognition system are the variation in the speaking styles. To deal with the variations in speaking styles, the variability in pronunciation must be handled with care. This paper shows the effects of including pronunciation variants for Punjabi language recognition.

The pronunciation modeling is to find out that at a lexical level, which variation was best modeled and which can be handled in a better way by the acoustic models as described by [1]. The variations in the segment of speech can be treated by the acoustic models using adaptation or training on the target speech. Some changes, i.e., insertions, deletions and dialects, and speaking styles, may be better taken into account more properly at the lexical level. Lexical modeling can handle much larger contexts as compared to acoustic modeling, allowing modeling of syllables and whole words or phrases [2]. But permitting much pronunciation variants in the lexicon may increase the probability of error.

A self-made lexicon will generally perform well as compared to the standard lexica in the recognition system, as discussed in The Linguistic Data Consortium, The Wall Street Journal speech database (2007) [3]. However, the availability of a handcrafted lexicon is costly and not much feasible. For the Punjabi language available handcrafted resources are very limited [4].

The evaluation of the use of pronunciation in variants both in the training of the acoustic models and in testing for the different speaking styles is done so that we can compare two models, the acoustic modeling and the lexical modeling of the variations.

Two methods do the training of acoustic models:

1. ‘Canonical Modelling’ is done by training using transcriptions based on a canonical lexicon

2. ‘Variant Modelling’ training is done using transcriptions based on a lexicon with variants

The first method will model all the variation using the acoustic models. And the second method will have low variations in the acoustic models, leaving more to lexical modeling. Monophonic and triphone models trained, i.e., context-independent, and context-dependent models trained.

Acoustic model adaptation is the way to handle variation that depends on speaking style using seed models which fit the speaking styles differently. The acoustic model adaptation is also dependent on the amount of available data.

Kessens, J.M. shows that in lexical task adaptation, the pronunciation probabilities improve the performance many times [1]. One way to perform a forced alignment on a development set and to estimate the probability by frequency counts as implemented [5].

II. Database Preparation

The audio database of Continuous Punjabi language made. The audio data recorded of hundred numbers of speakers out of which sixty speakers were male speakers and forty speakers were female speakers. The Malwai dialect of Punjabi language selected as the major regions of Punjab speaks malwai dialects only. The speakers selected mainly from Ludhiana and Hoshiarpur region of Punjab.

The audio recording was done on index mike. The mike kept at a distance of 2-3 mm from the speaker's mouth. Total of 100-hour audio data recording made on wave analyzer. The

Jyoti Guglani working as Assistant Professor in IMSEC, India (phone number: 9971715803; e-mail: jyotidhiman22@gmail.com).

A.N. Mishra working as Professor in Krishna Engineering College, India (e-mail: an.mishra58@gmail.com).

.

wave analyzer tool used for processing the speech signal. The background noise was removed using the software. The speech signal segmented for a speech recognition system. The segmented speech signal used for training and testing of the Automatic Speech Recognition system.

III. Methodology

In this paper, the training of a speech recognizer is done using the Kaldi toolkit. The MFCC and PLP feature extraction techniques used with HMM classification in Kaldi toolkit by Povey, D [6].

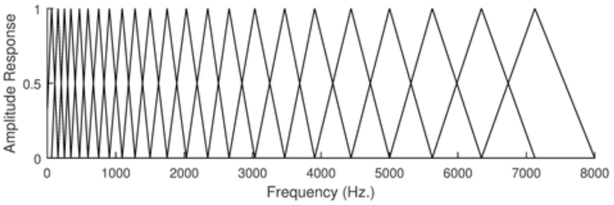

Fig 1. Mel-frequency scaled filterbank. The figure shows the amplitude response of the 23 filters of a Mel-scaled filterbank ranging from 0 to 8000 Hz.

The STFT is the fundamental section of speech analysis, but generally, DNNs do not act as input the spectrum directly. They use typically more compact features obtained as a result of STFT. With some modifications giving rise to the Mel frequency cepstral coefficients(MFCC), with specific focus to the influence of these transformations in the time-frequency resolution. The Mel Scale is shown in Fig. 1.

Deep neural networks (DNN) are machine learning tools with permit learning of complex non-linear multidimensional functions of a given input to reduced an error cost. A graphical example of a standard deep neural network is presented in Fig 2.

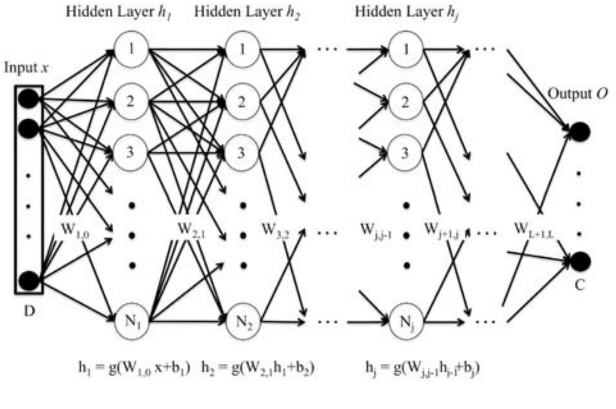

Fig 2. Deep Neural Network (DNN). A graphical representation of a standard feedforward DNN architecture. It is fed with an input vector *x* of dimension D which is transformed by the hidden layers h_j (composed of N_j hidden units) according to an activation function *g* and the parameters of the DNN (weight matrices *W* and bias vectors *b*)

As a result, a feedforward DNN used to perform a classification task might have the general structure. The structure is having; an input layer, fed with some input vectors representing the data; two or more hidden layers, with a transformation applied to the output of the previous layer, obtaining a higher level representation as we move away from the input layer; and an output layer, which computes the output of the DNN.

The HTK toolkit is also used for the speech recognition system for recognition by Woodland, P.C. and Steve Young [7], [8].

Kaldi toolkit is an open-source toolkit which is used for speech recognition and is written in C++ language. The advantage of Kaldi toolkit based speech recognition system is a fast real-time recognition described by D. Povey [6].

For training acoustic models using pronunciation variants, these variants were used to re-transcribe the training data with forced alignment. The most common scheme is to transcribe the training set first using monophone models and then use that transcription for the remaining of the training.

The pronunciation variants retranscribed the data with every increase of the number of Gaussians in the observation probability mixture. With every level of mixture components, we used Baum-Welch reestimation for iterations using the transcription from the previous level. Then the reiterated models have used to transcribe the data. The two different retranscription schemes were performed both for the monophone and triphone models.

For the canonically trained Hidden Markov Models, we performed four iterations for each level of mixture components for monophone and triphone models both as analyzed [10].

Retranscribing for each mixture update gave larger log likelihood data after each iteration compared to transcribing once. The difference in two was not much, which could be the result of the more iterations used in training. The results of recognition showed a less increase in performance for most conditions. The differences between the two schemes were statistically not significant. The results shown in the paper were derived using models with updated retranscription [11] .

IV. Experimental Results

The monophone models result in some improvements using lexical variants in the test, as shown in Figure 1. The use of variants in training and not in test shows much deterioration as compared to having variants in both training and testing for all speaking styles. For the triphone models improvement seen using variants is not as uniform as seen in Fig 3.

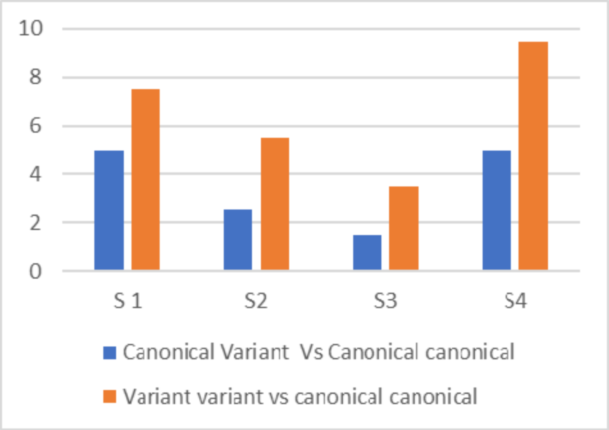

Fig 3: Relative WER improvement from canonical pronunciations in both training and test using monophonic

It is seen that there was a significant difference by using variants in training, but not in the test with all speech types and compared to other conditions.

Figure 3 shows the triphone model as compared to using canonical pronunciations in both training and testing as well as the variants in both training and testing condition. There is deterioration in performance for variants in acoustic model training and not in the testing lexicon.

The improvement from context-independent to context-dependent modeling gave a significant difference. There was no significant change for the canonical setup, and for the variant setup, there was a significant deterioration. The deterioration in the performance of the triphones model was worse than the canonical triphones models.

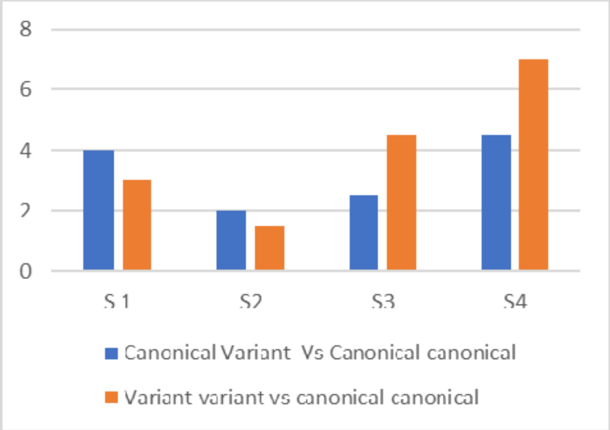

Fig 4: Relative WER improvement from canonical pronunciations in both training and test using triphones

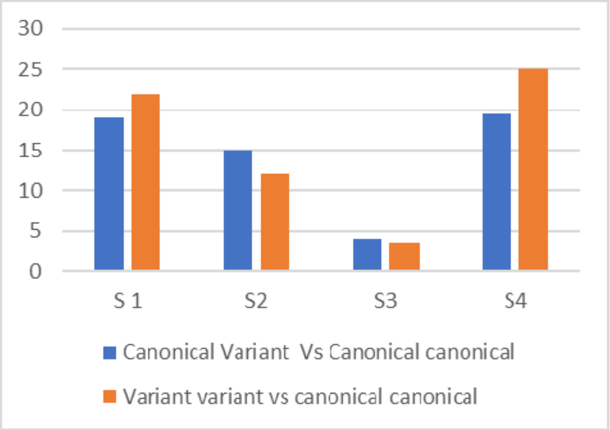

Fig 5. Relative WER deterioration from the matched conditions to the mismatch condition using triphones

The research shows similar results with other languages also and improving the recognition performance should be possible by careful selection of which pronunciation variants should be included [5]. To incorporate speaking style dependent lexical adaptation set is used by which pronunciation probabilities derived from forced alignment.

The observation of an experiment is the increase in system performance by taking care of pronunciation variants for speaking styles when using context-independent models. The increased modeling capacity of context-dependent models could handle the observed variation. We observed that the errors differed: About 20% of the errors were different when using variants compared to using only canonical pronunciations

The use of context-dependent models provides gain for all speaking styles besides non-native speech. The observations are the same as cited in a research paper by Van Compernolle [12]. Triphones trained in native speech are not suitable for modeling non-native speech.

The pronunciation variation modeling is a significant feature for the improvement of the performance of ASR systems. Holter, T. and Svedsen, T. (1999) give data-driven variant generation and lexicon optimization using an objective criterion is one such technique [13].

V. Conclusions

The improvements were observed for context-independent models where speaking styles shows much improvement. For the context-dependent models, the variants used for the speaking style of the acoustic model training set. For non-native speech, we saw not much improvement from context-independent to context-dependent modeling.

The probability of error rates was similar for the variant setup compared to the canonical setup, but there is a difference in errors. This shows that there is a selection potential in variants.

Two main concerns are important in pronunciation modeling: 1) speaker pronunciations, and 2) how to assess variations in pronunciation. To assess variants pronunciation, we require representative data. Various Methods based on rules of pronunciation rather than of directly on variants can generalize to pronunciations not present in the training data and will make it possible to assess these unseen pronunciation variants

REFERENCES

[1] Kessens, J. M., Wester, M., & Strik, H. (1999). Improving the performance of a Dutch CSR by modeling within-word and cross-word pronunciation variation. Speech Communication, 29(2-4), 193–207. doi:10.1016/s0167-6393(99)00048-5.

[2] Jurafsky, D., Ward, W., Zhang Banping, Herold, K., Yu Xiuyang, & Zhang Sen. (n.d.). What kind of pronunciation variation is hard for triphones to model? 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221). doi:10.1109/icassp.2001.940897D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O.

[3] The Wall Street Journal: Markets Data Center. (2007). Choice Reviews Online, 45(03), 45–1583–45–1583. doi:10.5860/choice.45-1583*(English) Lexicon Gregorianum Online.doi:10.1163/2214-8655_lgo_lgo_10_0222_eng*

[4] Besacier, L., Barnard, E., Karpov, A., & Schultz, T. (2014). Automatic speech recognition for under-resourced languages: A survey. Speech Communication, 56, 85–100. doi:10.1016/j.specom.2013.07.008

[5] McGraw, I., Badr, I., & Glass, J. R. (2013). Learning Lexicons From Speech Using a Pronunciation Mixture Model. IEEE Transactions on Audio, Speech, and Language Processing, 21(2), 357–366. doi:10.1109/tasl.2012.2226158

[6] Povey, D., Hannemann, M., Boulianne, G., Burget, L., Ghoshal, A., Janda, M.,Vu, N. T. (2012). Generating exact lattices in the WFST framework. 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). doi:10.1109/icassp.2012.6288848

[7] Woodland, P. C., Odell, J. J., Valtchev, V., & Young, S. J. (n.d.). Large vocabulary continuous speech recognition using HTK. Proceedings of ICASSP '94. IEEE International Conference on Acoustics, Speech and Signal Processing. doi:10.1109/icassp.1994.389562.

[8] Woodland, P. C., Leggetter, C. J., Odell, J. J., Valtchev, V., & Young, S. J. (n.d.). The 1994 HTK large vocabulary speech recognition system. 1995 International Conference on Acoustics, Speech, and Signal Processing. doi:10.1109/icassp.1995.479276The Linguistic Data Consortium, 1993

[9] Leggetter, C. J., & Woodland, P. C. (1995). Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. Computer Speech & Language, 9(2), 171–185. doi:10.1006/csla.1995.0010

[10] Karafiat, M., Burget, L., Matejka, P., Glembek, O., & Cernocky, J. (2011). iVector-based discriminative adaptation for automatic speech recognition. 2011 IEEE Workshop on Automatic Speech Recognition & Understanding. doi:10.1109/asru.2011.6163922

[11] Gillick, L., & Cox, S. J. (n.d.). Some statistical issues in the comparison of speech recognition algorithms. International Conference on Acoustics, Speech, and Signal Processing. doi:10.1109/icassp.1989.266481.

[12] Van Compernolle, D. (2001). Recognizing speech of goats, wolves, sheep and non-natives. Speech Communication, 35(1-2), 71–79. doi:10.1016/s0167-6393(00)00096-0.

[13] Holter, T., & Svendsen, T., (1999). Maximum likelihood modelling of pronunciation variation. Speech Communication, 29(2-4), 177–191. doi:10.1016/s0167-6393(99)00036-9

Jyoti Guglani*; India born in 1982. She received her Bachelor of Technology (B. Tech) degree in Electronics and Communication Engineering from U. P. Technical University in 2004, M. Tech in Digital Communication from Gautam Buddh Technical University. She is working as Assistant Professor in ECE department, Institute of Management Studies Engineering College, Ghaziabad since March 2014.Her research areas are Speech recognition, pattern recognition. Presently Assistant Professor in IMSEC, India*

A.N. Mishra*; India born in 1969. He received his Bachelor of Technology (B. Tech) degree in Electronics and Communication Engineering from Gulbarga University, Karnataka in 2000, M. Tech in Digital Communication from U. P. Technical University, Lucknow and Doctorate from Birla Institute of Technology, Mesra, Ranchi. He is working as Professor and Dean in ECE department, Krishna Engineering College, Ghaziabad since July 2013.His area of interest are speech recognition and synthesis. Presently Dean Student welfare and professor in KEC, India.*