MLLR ADAPTATION TECHNIQUES FOR PRONUNCIATION MODELING

Veera Venkataramani and William Byrne

Center for Language and Speech Processing
The Johns Hopkins University
Baltimore, MD 21218 U.S.A.
{*veera,byrne*}*@jhu.edu*

ABSTRACT

Multiple regression class MLLR transforms are investigated for use with pronunciation models that predict variation in the observed pronunciations given the phonetic context. Regression classes can be constructed so that MLLR transforms can be estimated and used to model specific acoustic changes associated with pronunciation variation. The effectiveness of this modeling approach is evaluated on the phonetically transcribed portion of the SWITCHBOARD conversational speech corpus.

1. INTRODUCTION

Pronunciation modeling for automatic speech recognition provides a mechanism by which ASR systems can be adapted to accented, spontaneous, or disfluent speech that is not well described by the canonical pronunciations found in dictionaries. One aspect of this modeling problem is to build predictive or descriptive models for phenomena of interest that can predict surfaceform pronunciation given the baseform or canonical pronunciation. Methods are available that can learn that the word IT can be pronounced as IH D rather than IH T or that the words GOING TO can be treated as a single entity pronounced G UH N AH [1].

Once these predictive models are trained they can be incorporated directly into an ASR system. An ASR system nominally consists of a language model $P(W)$; a dictionary that maps word sequences to baseform pronunciation sequences $P(B|W)$; and a set of acoustic models $P(A|B;\theta_B)$ that assign likelihood to the acoustic observations A given the baseforms B. The notation θ_B indicates that the acoustic model parameters were trained using baseform transcriptions of the acoustic training set. The pronunciation model is assumed to be available as a distribution $P(S|B)$ that maps baseforms to surfaceform sequences S; techniques that augment lexicons with frequent pronunciation alternatives or use decision trees to map baseform phone sequences to surfaceform sequences can be described in this way. The maximum likelihood decoder can be stated as

$$\underset{W,S,B}{\operatorname{argmax}} P(A|S,B)P(S|B)P(B|W)P(W) \qquad (1)$$

if appropriate conditional independent assumptions are made.

In addition to the pronunciation model, a particular form of acoustic model $P(A|S,B)$ is needed. A simple approximation is available as $P(A|S,B) \approx P(A|S;\theta_B)$, where acoustic models trained on baseform transcriptions are used directly with the surfaceform sequences produced by the pronunciation model. The ASR system is therefore able to produce word hypotheses based on pronunciations not present in the original dictionary. This straightforward approximation is especially effective because it allows the pronunciation model to be incorporated into the ASR system without retraining the ASR acoustic models. However since the acoustic models used in this approximation were trained on the baseform pronunciations, the recognition process is inevitably biased towards word hypotheses based on canonical pronunciations.

This observation leads to another aspect of the pronunciation modeling problem which is to incorporate models of pronunciation variability directly into acoustic modeling. Interestingly, it has been found that straightforward approaches to this problem often fail. One possible approach would be to train a pronunciation model; verify that it works well when used with a standard ASR system (by using the approach described in the previous paragraph); use this pronunciation model to retranscribe the acoustic training set to obtain a surface form transcription; retrain the acoustic models; and evaluate the new ASR system with the pronunciation model. This approach yields a set of models with parameters θ_S which can also be used to approximate $P(A|B,S)$ as $P(A|S;\theta_S)$. However, as has been discussed by Saraclar *et al.* [2], this can lead to degradation in ASR performance. Saraclar *et al.* conclude that it is incorrect to approximate $P(A|S,B)$ by either $P(A|S;\theta_B)$ or by $P(A|S,\theta_S)$. They demonstrate that when a base phone b is realized as a surfaceform s, the acoustic model should model it as such, *i.e.* it should model it not as an s but as a

particular variant of b. In other words, surfaceforms should not be modeled without consideration of the baseform from which they originate. In terms of modeling, $P(A|S,B)$ should retain dependencies on both baseform and surfaceform.

2. MLLR PRONUNCIATION MODELING

Our goal is to use acoustic model adaptation techniques to approximate the distribution $P(A|S,B)$ by transforming the parameters of the baseform ASR system $P(A|B;\theta_B)$. We assume that a surfaceform transcription of the acoustic training data is available, either from human annotators or through forced alignment using the pronunciation model and the acoustic models θ_B. We then align the surface annotations to the baseform transcriptions using a phonetic feature distance [1]. This symbol-to-symbol alignment allows us to construct a hybrid transcription for the training data: the original baseform sequence $\{b_j\}$ after alignment with the surface sequence $\{s_j\}$ becomes $\{b_j{:}s_j\}$. This hybrid transcription is used in supervised MLLR adaptation to estimate transforms $T_{S,B}$ that are applied to the parameters of the baseform models to make the approximation $P(A|S,B) \approx P(A|T_{S,B} \cdot \theta_B)$.

Phonetic transformation regression classes are used in modeling the potentially very large number of pairs *b:s*. Suppose an instance of the words SUPPOSE ITS with baseform transcription S AX P OW Z IH T S has the surfaceform annotation S IH P OW S IH D Z. After alignment the hybrid transcription becomes S:S AX:IH P:P OW:OH Z:S IH:IH T:D S:Z. A set of phonetic transformation regression classes could be defined as

$$T_{b,s} = \begin{cases} T_I & s = b, \text{ no change} \\ T_{voice+} & b \text{ unvoiced}, s \text{ voiced} \\ T_{nasal-} & b \text{ nasal}, s \text{ not nasal} \\ \dots & \end{cases}$$

In the examples given here, the transform T_{voice+} is trained on all data whose annotation indicates that an unvoiced baseform has changed to a voiced surfaceform, for example data labeled as S:Z or P:B.

Through the choice of regression classes we can adapt the models to the amount of available data or the expected phonetic variability. For example, it may be that consonants are observed to have little surfaceform variation, so regression classes might be constructed to describe only vowel variation. The classes need not be entirely complementary. For example, classes T_{voice+} and $T_{plosive:voice+}$ could coexist. Instances of both P:B and S:Z would be used to train the former, whereas instances of S:Z would not be used to train the latter. This allows a hierarchy of transforms that can be applied depending on the amount of training data available for each regression class.

The transform T_I associated with the *no change* phonetic transformation class is also estimated since the hybrid classes should be purer than the original baseform phonetic classes. For example, if an acoustic model was to be trained for P:P, all instances of P:T and other surfaceform variants would be excluded from training. This more homogeneous training set allows sharper acoustic models to be trained even for cases when no surfaceform variations are observed.

We note that adaptation techniques have been used before for pronunciation modeling. In dialect adaptation [3, 4] or in training a speaker dependent ASR system it is possible to use MAP or other acoustic adaptation techniques to refine the models to the new domain. It is assumed that sufficient data is available that the existing dictionary and model architecture are able to model the regular and consistent variations found in the data. However in previous work a predictive model of pronunciation change was not incorporated into acoustic model adaptation. The goal of this work is to explore the coupling of predictive pronunciation models with acoustic adaptation techniques.

3. PRONUNCIATION MODELING EXPERIMENTS

In the experiments we report here we focus on the prediction of surface pronunciations given the word sequence

$$\underset{S,B}{\operatorname{argmax}} P(A|S,B)P(S|B)P(B|W). \qquad (2)$$

This paradigm isolates the prediction of phonetic variation from the larger problem of incorporating pronunciation models into an ASR system with the goal of reducing word error rate. Performance is measured relative to phonetic transcriptions provided by expert phoneticians. We use the test and training set definitions and evaluation procedures established for the phonetic evaluation component of the 2000 Large Vocabulary Conversational Speech Recognition evaluation [5] that makes use of the ICSI phonetically transcribed SWITCHBOARD collection [6].

Baseform acoustic models $P(A|B;\theta_B)$ consisting of 48 monophone models were trained as in the JHU 2000 phonetic evaluation system [5]. The models were estimated on the training portion of the ICSI data using the phonetic transcription obtained from the lexicon; we note that monophone models have been found to be better for the prediction of surface variation than triphones. Each model was a three state left-to-right HMM with an 8 mixture, diagonal covariance Gaussian output density trained using HTK [7]. Surfaceform monophone acoustic models $P(A|B;\theta_S)$ with the same structure were also trained on the same data using the ICSI surfaceform transcriptions.

The decision tree pronunciation model [1] used to approximate $P(S|B)$ was based on the JHU 2000 phonetic evaluation system [5]. The models were trained on the train-

ing portion of the training set and incorporated only intra-word phonetic context; cross-word phonetic context was not used.

The availability of the surfaceform and baseform acoustic models allow us to approximate $P(A|S,B)$ in Equation 2 as either $P(A|\theta_S)$ or $P(A|\theta_B)$. The pronunciation model was applied to the test set word transcriptions to generate lattices of pronunciation alternatives for the test set utterances. As reported by Saraclar *el al.* [8], the surfaceform trained acoustic models gave the best phone error rate relative to the reference ICSI test set transcriptions.

Acoustic Model	Phone Error Rate (%)
Baseform	21.75
Surfaceform	20.50

Table 1: Baseform and Surface Acoustic Model Performance.

To train the MLLR transforms to be used as pronunciation models a surfaceform-tagged baseform transcription of the training set was produced by a symbolic alignment of the baseform transcriptions to the surfaceform transcriptions using phonetic feature distances [1]. For given sets of regression classes, each regression class transform was trained with six iterations of MLLR. Only mean transforms were estimated; variances were not adapted.

Tagged lattices were created from the lattices of pronunciation alternatives by tagging each surfaceform lattice link by the baseform phone from which it originated. Deletion arcs were left untouched. Only two instances of insertion were modeled: *en* $\rightarrow$ *en n* and *el* $\rightarrow$ *el l*. After MLLR transform estimation, decoding was done on the tagged-test-set lattices. Transforms were then applied to the baseform acoustic models according to the regression class of each tagged lattice link and Viterbi rescoring of the lattice was performed to find a string of tagged phones; the surfaceform tag sequence was taken as the hypothesis surfaceform pronunciation.

Class	Phones	Class	Phones
fl	AE	ch	IX UX
fml	EH	bl	AA AY AW
fmh	IH EY ER	bml	AO OW OY
fh	IY Y	bmh	UH
cml	AH AHI	bh	W UW
cmh	AX EL EN AX EM	v	all vowels

Table 2: Base Acoustic Classes Used to Construct Phonetic Transformation Regression Classes. Classes are based on vowel manner and place of articulation: front, central, back, high, middle, low.

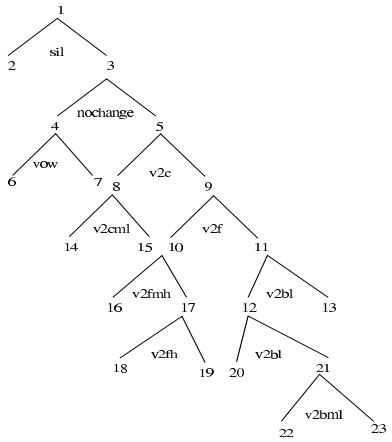

Figure 1: Phonetic Transformation Regression Tree. Identitiy transformations for silence, vowels, and consonants are defined at nodes 2, 6, and 7.

We report results using phonetic transformation regression classes based on the vowel groupings listed in Table 2. Regression classes based on consonant changes yielded very little improvement when used alone and had little effect when used along with vowel change regression classes. This is consistent with the observed behavior of both the baseform and surfaceform phone recognition systems which recognize consonants more reliably than vowels.

Figure 1 shows the regression tree with the phonetic transformation classes used in these MLLR pronunciation modeling experiments. The $v2f$ (*vowel2front*) label, for example, associated with node 10 specifies a regression class for baseform - surfaceform pairs b:s such that b is a vowel and s is a front vowel. The regression tree node indices also give the order in which the regression classes were created.

For purposes of comparison, we constructed regression trees using the routines provided by the HTK 3.0 Toolkit [7]. These routines create regression classes based on inter-Gaussian distances, but without consideration of phonetic similarity; relatively little improvement in Phone Error Rate (PER) over the baseline was observed using these classes. Performance was also found to be sensitive to the choice of regression classes. For example, adding the class *v2c* improves PER, while adding a class that allows only changes in location, *i.e.* non-central vowels realized as central vowels, does not improve PER.

Table 3 presents recognition results showing that phone error rate improves as phonetic transformation classes are added. As more classes are added performance approaches that of acoustic models trained directly on surface form transcriptions. Clearly, enough classes can be added so that each state, and eventually each Gaussian component, will be trained individually which would produce the surfaceform acoustic models. In these simple experiments we do not

Regression Classes	Class Added	Phone Error Rate (%)
Baseform Acoustic Models		21.75
1	global	21.70
2	silence	21.72
3	no change	21.49
4	cv	21.38
5	v2c	21.32
6	v2f	21.16
7	v2b	21.06
8	v2cml	20.99
9	v2fmh	20.94
10	v2fh	20.91
11	v2bl	20.91
12	v2bml	20.86
Surfaceform Acoustic Models		20.50

Table 3: Pronunciation Modeling Performance Showing Phone Error Rate Reduction as Phonetic Transformation Regression Classes are Introduced.

expect an improvement over the surfaceform models since the corpus is fairly homogeneous and there is sufficient data to train each individual surfaceform model. These experiments demonstrate however that phonetic transformations defined in terms of broad acoustic classes are able to capture nearly all of the predictive gains that can be obtained using detailed acoustic models trained on surfaceform transcriptions.

4. CONCLUSION

Phonetic transformation regression classes are introduced to model acoustic change associated with pronunciation variation. Ideally the application of this approach will allow a hierarchy of phonetic transformation classes to be defined in which individual baseform-surfaceform pairs are assigned to the most appropriate class based on acoustic similarity. Automatic techniques for constructing regression classes for MLLR could also be used to develop hierarchies in an unsupervised manner. Ultimately it is hoped that these techniques will allow for the development of detailed transformation procedures that improve the adaptation of ASR systems to new speakers and dialects through the tighter coupling of acoustic adaptation and detailed models of phonetic variation.

Acknowledgement. We thank Michael Riley and Murat Saraclar of AT&T for use of the Large Vocabulary Decoder and for assistance with pronunciation modeling tools based on the FSM toolkit.

5. REFERENCES

[1] M. Riley and A. Ljolje. Automatic generation of detailed pronunciation lexicons. *Automatic Speech and Speaker Recognition : Advanced Topics*, Kluwer Academic, 1995.

[2] M. Saraclar. and S. Khudanpur. Pronunciation ambiguity vs pronunciation variability in speech recognition. *Proc. Eurospeech*, pp. 515-518, Budapest, 1999.

[3] V. Digalakis *et al.* Development of dialect-specific speech recognizers using adaptation methods *Proceedings of the IEEE ICASSP*, pp 1455-1458, 1997.

[4] J. J. Humphries, and P. C. Woodland. The use of accent-specific pronunciation dictionaries in acoustic model training *Proceedings of the IEEE ICASSP* pp 317-320, 1998.

[5] 2000 NIST Evaluation Plan for Recognition of Conversational Speech over the Telephone `www.nist.gov/speech/tests/ctr/h5_2000/h5-2000-v1.3.htm`

[6] S. Greenberg. The Switchboard Transcription Project. in Research Report 24, 1996 Large Vocabulary Continuous Speech Recognition Summer Research Workshop Technical Report Series. Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD. `www.icsi.berkeley.edu/real/stp`

[7] S. Young *et al.* The HTK Book (Version 3.0), July 2000.

[8] M. Saraclar. Pronunciation Modeling, Ph. D. Thesis, The Johns Hopkins University, June 2000.

[9] W. Byrne, *et al*. Pronunciation modeling using a hand-labelled corpus for conversational speech recognition. *Proc. ICASSP*, pp 313-316, 1998.

[10] M. Riley, *et al.* Stochastic pronunciation modeling from hand-labelled phonetic corpora. *Speech Communication*, 29(2-4):209–224, November 1999.

[11] C. J. Legetter and P. C. Woodland. Maximum Likelihood linear regression for speaker adaptation of continuous density HMM's *Computer, Speech and Language*, vol. 9, pp. 171-186, 1995.