Modeling pronunciation variation using artificial neural networks for English spontaneous speech

Ken Chen, Mark Hasegawa-Johnson

Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign
{kenchen, jhasegaw}@uiuc.edu

Abstract

Pronunciation variation in conversational speech has caused significant amount of word errors in large vocabulary automatic speech recognition. Rule-based approaches and decision-tree based approaches have been previously proposed to model pronunciation variation. In this paper, we report our work on modeling pronunciation variation using artificial neural networks (ANN). The results we achieved are significantly better than previously published ones on two different corpora, indicating that ANN may be better suited for modeling pronunciation variation than other statistical models that have been previously investigated. Our experiments indicate that binary distinctive features can be used to effectively represent the phonological context. We also find that including pitch accent feature in input improves the prediction of pronunciation variation on a ToBI-labeled subset of the Switchboard corpus.

1. Introduction

Contemporary large vocabulary automatic speech recognizers (ASRs) require pronunciation dictionaries to integrate the recognition effort at acoustic and language levels. Most pronunciation dictionaries used in current ASRs contain only a few alternative pronunciations for most words. This has brought at least two problems. In recognition, pronunciation variation creates confusion to the decoding algorithm. For example, the word “them” can be pronounce as /dh ax n/ in some contexts, bearing no acoustic difference from the word “than”. In iterative acoustic training, where phonetic transcriptions are generated from word transcriptions and pronunciation dictionaries via forced alignment and used to retrain the acoustic models, the amount of modeled pronunciation variation in the dictionaries will significantly affect the accuracy of the resulting acoustic models (and the resulting phonetic transcriptions). Due to the above reasons, seeking models that accurately account for pronunciation variation has been an active research topic during the past several decades.

Oshika et al.[1] formulated common types of pronunciation variation in English as a series of rules including vowel reduction/deletion, alveolar flapping, homorganic stop deletion, geminate reduction and etc. In the literature, various knowledge-based approaches and data-driven approaches have been proposed [2]. Among them, the decision-tree based approaches have received a considerable amount of attention and have created successful results on standard English corpora [3, 4, 5]. Although pronunciation variation occurs both word-internally and across word boundaries (with possibly different characteristics), phone-based models that predict the realized phones (the surface form) from a window of canonical phonemes (the citation form) are used for both types of variation. Syllable-based and word-based models were investigated by Fosler-Lussier[5] and were shown to capture many of the coordinated phone pronunciation variations not handled by independent phone models.

It has been shown [4, 5] that the mapping from canonical phones to surface phones is dynamic, in the sense that realization of the current phone depends on realization of previous phones. It has also been demonstrated that auxiliary factors such as stress, syllabification, syntax and prosody may have an effect on pronunciation. We are particularly interested in the effect of prosody on pronunciation. Words bearing pitch accent are usually “hyper-articulated,” and have been reported to suffer less co-articulation than other words. Discrete pitch accent tags have not, we believe, previously been used in pronunciation models, but auxiliary features related to pitch accent have been shown to improve pronunciation models. Fosler [5] achieved better performance by including speaking rate in input feature vectors. Ostendorf et al. [6] reported a small performance improvement when word level syntactic features (part-of-speech tags) and acoustic-prosodic features (F0, duration and energy) are included in input feature vectors.

Neural networks have been proposed to model pronunciation variation in Japanese [7]. Our approach may be considered an extension of the work in [7], but with three extensions. First, we include a number of input features that have proven useful in tree-based pronunciation models, including syntactic and dynamic features;

because of the similarity in our methods, we are able to compare our results directly to those achieved by a widely cited and carefully designed tree-based model [4]. Second, we experiment with the relative merits of distinctive feature encoding of phones, in place of the phoneme indicator functions used by [7]. Third, we consider the use of explicit prosodic tags (specifically, a binary tag representing presence vs. absence of pitch accent) as an input feature for dynamic pronunciation modeling.

2. Methods

2.1. The models

The problem under investigation in this study is to predict the realized phonetic sequence $\hat{Q} = [\hat{q}_0, \ldots, \hat{q}_N]$ (obtained from acoustic signal by phoneticians or via automatic methods) given the canonical phoneme sequence $Q = [q_0, \ldots, q_M]$ (created by concatenating the pronunciation of individual words contained in the lexicon), where the phonetic variables $\hat{q}_i$ and q_i take values on the same phoneme set Ω. In order to incorporate the prediction scores into the probabilistic framework of ASRs, probability density functions (PDFs) are usually used as predictors, and an auxiliary vector sequence $A = [\vec{a}_0, \ldots, \vec{a}_M]$ containing high level information that has known effects on the prediction are included as conditioning factors: $\hat{p}(\hat{q}_i | q_0, \ldots, q_M, \vec{a}_0, \ldots, \vec{a}_M, \hat{q}_0, \ldots, \hat{q}_{i-1})$. Obviously, not all these conditioning factors are relevant. Therefore, a function $\phi(q_0, \ldots, q_M, \vec{a}_0, \ldots, \vec{a}_M, \hat{q}_0, \ldots, \hat{q}_{i-1})$ can be used to select the relevant factors. Usually the factors that affect the prediction of $\hat{q}_i$ are assumed to be localized within a small window of L phonemes centered around q_i, and the previous realized phone $\hat{q}_{i-1}$ is also selected, making this system first-order Markov [3]:

$$\begin{aligned} &\hat{p}(\hat{q}_i | \phi(q_0, \ldots, q_M, \vec{a}_0, \ldots, \vec{a}_M, \hat{q}_0, \ldots, \hat{q}_{i-1})) \\ &\quad \approx \quad \hat{p}(\hat{q}_i | q_{i-\delta}, \ldots, q_{i+\delta}, \vec{a}_{i-\delta}, \ldots, \vec{a}_{i+\delta}, \hat{q}_{i-1}), \end{aligned} \tag{1}$$

where $\delta = (L-1)/2$ and L normally equals 3 or 5. The auxiliary vector $\vec{a}_i$ is composed of several variables describing high level linguistic information.

2.2. The transcriptions and the dictionary

In order to train the pronunciation model, realized phone transcriptions (manually transcribed) are automatically aligned with a baseline phoneme transcription, constructed by concatenating canonical pronunciations from the Pronlex dictionary. Automatic alignment minimizes a metric similar to Levenshtein distance, but with a substition cost sensitive to the number of distinctive features that differ between baseline phoneme and realized phone. An example of the alignment is given below:

```
/ae n d w ah t  y uw k ae n t t ey k/
/eh n # w ax ch # uw k ae n # t ey k/
```

for the utterance "and what you can't take" from Switchboard. The first row in this example is the phoneme transcription and the second row the hand-labeled phone transcription. Notice the alveolar stop deletions (/ae n d/ → /eh n/), vowel reduction (/ah/ → /ax/) and palatalization (/t y/ → /ch/). Symbol "#" is used to represent deletions and insertions.

To be consistent with previous published works [4], Pronlex, a dictionary distributed by LDC, is used to generate the phoneme transcription. Pronlex contains a set of about 48 phonemes similar to those used in TIMIT.

The complete set of aligned transcriptions is available at http://www.ifp.uiuc.edu/speech.

2.3. Performance Measure

In this paper, we measure the quality of our model using cross entropy:

$$H(T) = -\frac{1}{R} \sum_{i=1}^{R} \log \hat{p}(\hat{q}_i | q_{i-\delta}, \ldots, q_{i+\delta}, \vec{a}_{i-\delta}, \ldots, \vec{a}_{i+\delta}, \hat{q}_{i-1}), \tag{2}$$

where T is a test set of aligned transcriptions that contain R phones. The more accurate the PDFs are, the larger the probability scores in average, and the smaller $H(T)$.

In computing (2), Riley et al. [4] exclude the worst 10% log probability scores in the summation. We apply the same strategy in this paper so that our results will be comparable with theirs on the same databases.

2.4. The ANN architecture and the input features

The ANN used in this study is a multi-layer perceptron (MLP) trained with error back propagation. The number of output nodes is equal to the size of the phoneme set plus an additional node "null" used for phone deletion. We currently do not handle insertions in this system as the number of insertions is small in hand-transcribed phonetic corpora.

We compare two possible phoneme encodings: an encoding using binary indicator functions (as in [7]), and an encoding using distinctive features. Distinctive feature representations may be either binary or multi-valued; for example, [4] encodes each phoneme using four distinctive features (consonant-manner, consonant-place, vowel-manner, vowel-place), each of which takes 8 to 10 different categorical values. Few ASR models use binary distinctive features, but theoretical phonological models often do; for example, Keyser and Stevens [8] represent coarticulation and reduction as the spreading or deletion of conditionally independent binary distinctive features. In our study, each phoneme is described using a vector of 15 binary distinctive features including 7 features for vowels: "high", "low", "back", "diphthong", "tense", "reduced" and "rounding", 7 features for the manner, place and voicing of consonants: "sonorant", "continuant", "syllabic", "blade", "anterior", "distributed" and "spread glottis", and a feature "vocalic" that distinguishes vowel vectors from consonant vectors.

For each phone, five auxiliary features are encoded: $\vec{a}_i = [b, l, s, f, p]$, where b encodes phone position relative to the closest word boundary, l phone position in the syllable, s lexical stress, f function word versus content word, and p pitch-accented vs. unaccented. These features are all binary except b, which is integer-valued. We divide b by a constant such that it ranges between 0 and 1. Function words are chosen based on their part-of-speech and word frequency. Feature p is only used in our experiments on the prosodically transcribed corpus.

3. The Corpora

Our experiments are conducted on three English corpora: TIMIT, ICSI (the ICSI spontaneous-speech phonetically transcribed corpus [9]), and ICSI_Prosody (a subset of the ICSI corpus that has been prosodically transcribed at the University of Illinois [10]). In TIMIT, all the si and sx sentences from the train and test directories are used for training and testing. Altogether, there are around 134,000 phones. The 61 phonemes in the TIMIT phoneme set are collapsed into a 42-phoneme set in order to be consistent with the Pronlex phoneme set (a few phonemes in Pronlex are also merged). The ICSI corpus contains around 96000 phones that were originally hand-transcribed using a very detailed phonetic label set; these transcriptions are also collapsed into the 42-phoneme set. ICSI_Prosody has the same phone transcriptions as the rest of the ICSI corpus, but also carries ToBI [11] prosodic transcriptions including the presence vs. absence of pitch accent. ICSI_Prosody is the result of a currently ongoing transcription effort. Current ICSI_Prosody transcriptions contain only about 5,000 phones, therefore we use all available data in this corpus for training, and evaluate the effects of prosody by comparing the training performance with vs. without prosodic features.

4. Results

The first experiment compares the performance of our ANN based models with that of the decision-tree based models reported in [4]. We believe that our results are comparable with theirs because we used the same corpora, same dictionary and same performance measure as theirs, but it is possible that minor differences in implementation details exist. Table 1 compares the results reported in [4] and the results of our best performing networks.

Table 1 indicates that ANN based models reduce the cross-entropy by 71.2% on TIMIT and 51.5% on ICSI, which are significantly higher than reductions obtained from the decision tree models (51% on TIMIT, 30% on ICSI). The baseline cross-entropy before training (the numbers on the left of →) is computed by replacing the log probability in equation (2) with the unigram probabilities $p(\hat{q}_i|q_i)$ estimated from the training data. The best performing MLP in our experiment contains a single hidden layer with around 40 nodes.

	Decision Trees		**ANNs**	
	Entropy	**%**	**Entropy**	**%**
TIMIT	0.34→0.17	51.0	0.358→0.103	71.2
ICSI	0.72→0.50	30.0	0.835→0.405	51.5

Table 1: The absolute and the percent (%) reduction of cross entropy (bits) for decision-tree based models and ANN based models on TIMIT and ICSI. Decision tree results are as reported in [4].

	Coding	L	H	**Cross Entropy**	**%**
IF3	Indicator	3	28	0.358→0.106	70.4
IF5	Indicator	5	20	0.358→0.105	70.7
DF3	Dist. Feat.	3	57	0.358→0.103	71.2
DF5	Dist. Feat.	5	44	0.358→0.119	66.8

Table 2: The absolute and the percent (%) reduction of cross entropy (bits) on TIMIT for model IF3, IF5, DF3, DF5 with difference in the coding scheme (indicator function (IF) or distinctive feature (DF)), window size L, and number of hidden nodes (H).

Our second experiment on TIMIT compares the performance of different network configurations. We are interested in finding out whether a larger window size ($L = 5$) gives better performance than a smaller window size ($L = 3$). We are also interested to know if a distinctive feature (DF) based encoding scheme produces better performance than an indicator function (IF) based scheme. For fair comparison, we adjust the number of hidden nodes such that all the models have approximately equal number of parameters. The previous realized phone $\hat{q}_{i-1}$ is included in the input vector in all models. The results of this experiment are listed in table 2.

As shown in table 2, increasing the window size L does not improve the performance (and actually hurts the DF based models). DF based systems yields slightly better performance than IF based systems when $L = 3$. This result suggests that relevant information is mostly contained in a 3 phoneme window. Increasing window size apparently fragments the training data and reduces the ability of models to generalize from training to test data. DF encoded phonological context is at least as effective as IF encoded phonological context given a small window size.

The third experiment tests the hypothesis that pitch accent affects the phone realization probability. The binary pitch accent feature p (accented vs. unaccented syllable) is appended to the input vector, with values extracted from the ToBI labels. The performance on

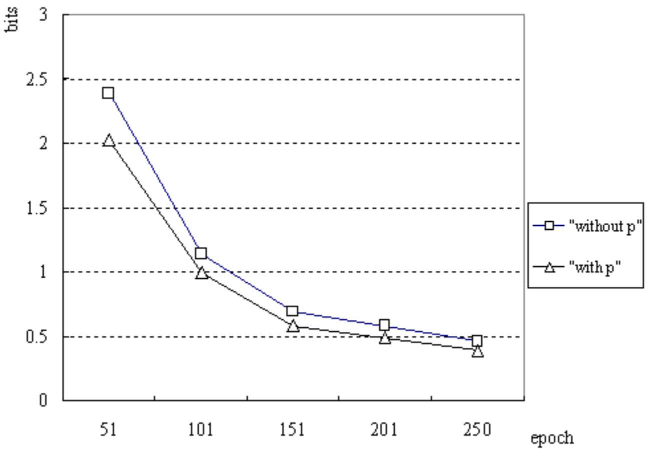

Figure 1: *The averaged entropy on training set as a function of training iterations with and without pitch accent feature p.*

ICSI_Prosody is compared in terms of the entropy, the mean squared error and the prediction accuracy on the training set after networks were trained from the same initialization for 250 iterations. Consistent improvement is found with feature p included under most of the conditions. We plot the averaged entropy (averaged over all conditions) with and without p on fig. 1. Obviously, smaller averaged entropy is achieved when p is included in the input vector.

5. Conclusions

In this paper, we report our work on modeling pronunciation variation using artificial neural networks (ANN). Results indicate that ANNs may be well suited for modeling pronunciation variation. A binary distinctive feature encoding of input phones performs slightly but significantly better than indicator features over a three-phone input window, but suffers from overtraining when used with a five-phone window. We also find that including pitch accent feature in input improves the prediction of pronunciation variation on a subset of ToBI labeled switchboard corpus.

6. Acknowledgement

The authors would like to thank S. Chavarria, T. J. Yoon and J. Cole for providing the prosody labels and S. Borys in preparing phonetic transcriptions and dictionaries.

7. References

[1] Oshika, B. T., Zue, V. W., Weeks, R. V., Neu, H., and Aurbach, J., “The Role of Phonological Rules in Speech Understanding Research”, IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 23, pp. 104-112, 1975.

[2] Strik, H. and Cucchiarini, C., “Modelling pronunciation variation for ASR: Overview and comparison of methods”, in Proceedings of the ETRW workshop on Modelling Pronunciation Variation for ASR, Rolduc, The Netherlands, 1998.

[3] Riley, M. and Ljolje, A., “Automatic generation of detailed pronunciation lexicons”, in Automatic Speech and Speaker Recognition: Advanced Topics, Kluwer, Boston, 1995.

[4] Riley, M., Byrne, W., Finke, M., Khudanpur, S., Ljolje, A., McDonough, J., Nock, H., Saraclar, M., Wooters, C., and Zavaliagkos, G., “Stochastic pronunciation modelling from hand-labelled phonetic corpora”, Speech Communication, vol. 29, pp. 209-224, 1999.

[5] Fosler-Lussier, E., “Contextual word and syllable pronunciation models”, in Proceedings of the IEEE workshop on Automatic Speech Recognition and Understanding, Keystone, Colorado, U.S.A, 1999.

[6] Ostendorf, M., Shafran, I., and Bates, R., “Prosody models for conversational speech recognition”, in Proc. for 2002 Plenary Meeting and Symposium on Prosody and Speech Processing, Tokyo, Japan, 2002.

[7] Fukada, T., Yoshimura, T., and Sagisaka, Y., “Automatic generation of multiple pronunciations based on neural networks”, Speech Communication, vol. 27, pp. 63-73, 1999.

[8] Keyser, S. J. and Stevens, K. N., “Feature geometry and the vocal tract”, Phonology, vol. 11, pp. 207-236, 1994.

[9] Greenberg, S., “The switchboard transcription project”, in 1996 LVCSR Summer Workshop Technical Reports, http://www.icsi.berkeley.edu/real/stp/.

[10] Chavarria, S., Yoon, T. J., and Cole, J., “Acoustic differentiation of ip and IP boundary levels: Comparison of L- and L-L% in the Switchboard corpus”, in Proc. ICSA International Conference on Speech Prosody 2004, Nara, Japan, March 2004.

[11] Beckman, M. E. and Elam, G. A., “Guidelines for ToBI labelling,” http://www.ling.ohio-state.edu/research/phonetics/E_ToBI/singer_tobi.html, 1994.