A STATISTICAL LEXICON BASED ON HMMS

Rainer Gruhn, Satoshi Nakamura

ATR Spoken Language Translation Res. Labs.
2–2–2 Hikaridai, Keihanna Gakkentoshi, Kyoto 619-0288, Japan
rainer.gruhn@atr.jp

ABSTRACT

This paper introduces a novel approach towards pronunciation modeling for pronunciation rescoring. Rather than explicitly representing pronunciation variations, a discrete HMM is provided for each word, modeling seen and allowing unseen pronunciation variations. Phone substitutions, deletions and insertions are equally covered. The approach is evaluated on non-native speakers speech recognition task.

1. INTRODUCTION

A lot of work has been reported about pronunciation modeling [1]. Many approaches follow the similar basic scheme of comparing manual or automatically generated phoneme transcriptions to some baseline transcription. Variation information can be extracted from the differences. Typically it is represented in the form of rules, which can be weighted based on occurence frequency, likelihood, confusability or other measures. These rules are applied to a baseline lexicon in order to generate some adapted lexicon or to optimize an acoustic model. Unfortunately this approach usually brings only little improvement.

In this research, we suggest a new data-driven approach to deal with pronunciation variations. It is based on word-level pronunciation HMMs, which are applied to rescore n-best hypotheses. Our target is to improve the performance of a continuous speech recognition system on a challenging speaker group such as non-native speakers.

Similar to the standard approach, we generate a phonetic transcription with phoneme recognizer. These phoneme sequences are used as training data for discrete word HMMs; one HMM for each word. There is no attempt to explicitly represent the phoneme variations. Even variations unseen in the training data are allowed, as a certain floor probability exists for all possible phoneme sequences for each word. The HMM training process will implicitly take care of all variation- and likelihood issues, unlike in other approaches, e.g. rule firing frequencies do not have to be calculated.

2. WORD HMMS

As illustrated in Fig. 1, two levels of HMM-based recognition are involved in this approach:

- Acoustic level: phoneme recognition to generate the phoneme sequence S_i from the acoustic features O_i
- Phoneme label level: For training, the phoneme sequences S_i are considered as input. For all words, a discrete word HMM is trained on all instances of that word in the training data. The models are applied for rescoring, generating a pronunciation score given the observed phoneme sequence S_i and the word sequence.

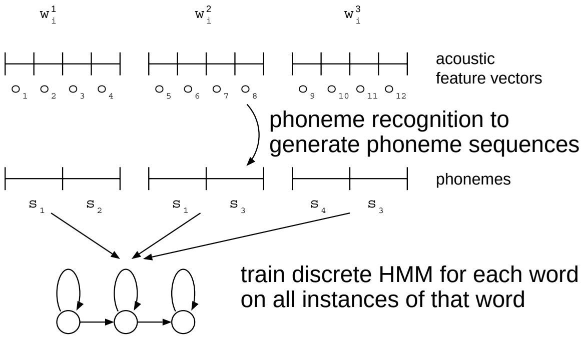
w_i^1
w_i^2
w_i^3
acoustic feature vectors
o_1 o_2 o_3 o_4 o_5 o_6 o_7 o_8 o_9 o_{10} o_{11} o_{12}
phoneme recognition to generate phoneme sequences
phonemes
s_1 s_2 s_1 s_3 s_4 s_3
train discrete HMM for each word on all instances of that word

Figure 1: *Two layers of HMMs are required to generate pronunciation variants and their likelihoods: an acoustic level for phoneme recognition and the phoneme label level for word model training.*

The first step requires a standard HMM acoustic model, and preferably some phoneme bigram language model as phonotactic constraint. The continuous training speech data is segmented to word chunks based on time information generated by viterbi alignment. Acoustic feature vectors are decoded to an 1-best sequence of phonemes.

For each word in the vocabulary, one discrete untied HMM is generated. Figure 2 shows an example for the word “and”.

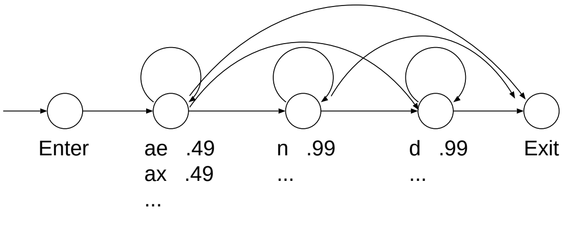

Figure 2: *An example discrete word HMM for the word “and”, initialized with two pronunciation variations for the first phoneme.*

The models are initialized on the phoneme sequence in some baseline pronunciation lexicon. The number of states for a word model is set to be the number of phonemes in the baseline pronunciation, plus enter and exit states. Each state has a discrete probability distribution of all phonemes, giving the baseline phoneme a high probability and all other phonemes some low but non-zero value. Forward transition between all states is allowed, with initial transition probabilities favouring a path that hits each state once.

The probability distribution as well as the transition probabilities are reestimated on the phoneme sequences of the training data. For each word, all instances in the training data are collected and analyzed. The number of states of each word model remains static. Phoneme deletions are covered by state skip transitions, phoneme insertions are modeled by state self-loop transitions.

Data sparseness is a common problem for automatically trained pronunciation modeling algorithms. In this approach, pronunciations for words that do appear sufficiently frequent in the training data, the pronunciations are generated in a data-driven manner. For rare words, the algorithm falls back on baseline phoneme sequences from a given lexicon. This combination should make it more robust than for example an application of phoneme confusion rules on a lexicon (as e.g. in [2]) could.

3. EXPERIMENTS

3.1. Phoneme recognition

For evaluation, we used a non-native database collected at ATR and consisting of 11 Japanese speakers of English. About 12 minutes of read speech are available per speaker, which was divided into ten minutes for training and two minutes as test set. The task domnain is hotel reservation.

The non-native training data set is segmented into single words based on time information aquired by viterbi alignment. On these word chunks, phoneme recognition is performed. To archieve higher phoneme recognition accuracy than with monophones, a right-context biphone model is applied. In the resulting phoneme string, the context is not considered, though. The phoneme recognition accuracy for the non-native task is 34.68% relative to the canonic transcription. The biphone acoustic model in this experiment is trained on the Wall Street Journal (WSJ) read speech corpus [3] The phoneme set consists of 43 phonemes plus silence. In the second level of processing, the rescoring, occurences of silence are ignored. The HTK toolkit [4] is used for all training and decoding steps.

3.2. Word HMM initialization

The discrete probability distribution for each state is initialized depending on the “correct” phoneme sequence(s) as given in the lexicon. The correct phoneme has a probability of 0.99; if more than one pronunciation variant is included in the lexicon, the variations all have the same probability. All other phonemes are assigned some non-zero probability.

The transition probabilities depend on the number of succeeding phonemes in the baseline lexicon. The probability to skip k phonemes is initialized to 0.05^k. Insertions are allowed with a chance of 0.05. The transition to the next state therefore has a probability of slightly below 0.9.

3.3. Rescoring

The HMM pronunciation models are applied in the form of rescoring the n-best decoding result. On an utterance in the test data, both a 1-best phoneme recognition and a standard n-best recognition (on word level) is performed. For each of the n-best sequences, we apply a forced alignment using the discrete pronunciation models, the phoneme sequence as input features and the word sequence as labels. The resulting score is the pronunciation score.

This pronunciation score is combined with the weighted language model score for this hypothesis. The hypothesis archieving the highest total score among the n-best is selected as correct. Figure 3 shows the performance for various language model weights. The best performance is 29.04% word error rate (WER) compared to baseline performance in this experiment of 32.54%.

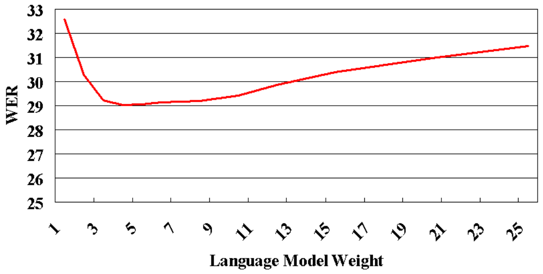

Figure 3: *Word error rate for rescoring of n-best based on pronunciation score combined with weighted language model scores.*

4. CONCLUSION

Word error rate could be improved by a relative 10.8% with pronunciation rescoring, showing the effectiveness of the approach for non-native speech. The full strength of the approach may not be achieved in this evaluation because of lack of non-native training data, which frequently forces word models to default to the standard pronunciations. Also, considering the acoustic score together with pronunciation and language model score could be a helpful extension.

5. ACKNOWLEDGEMENT

The research reported here was supported in part by a contract with the Telecommunications Advancement Organization of Japan entitled, ”A study of speech dialogue translation technology based on a large corpus”.

6. REFERENCES

[1] Helmer Strik and Catia Cucchiarini, “Modeling pronunciation variation for ASR: A survey of the literature,” *Speech Communication*, vol. 29, pp. 225–246, 1999.

[2] Rainer Gruhn, Konstantin Markov, and Satoshi Nakamura, “Probability sustaining phoneme substitution for non-native speech recognition,” in *Proc. Acoust. Soc. Jap.*, Fall 2002, pp. 195–196.

[3] D.B. Paul and J.M.Baker, “The design for the wall street journal based CSR corpus,” in *Proc. DARPA Workshop*, Pacific Grove, CA, 1992, pp. 357–362.

[4] P. Woodland and S. Young, “The HTK tied-state continuous speech recognizer,” in *Proc. EuroSpeech*, 1993, pp. 2207–2210.