WORD-LEVEL RATE OF SPEECH MODELING USING RATE-SPECIFIC PHONES AND PRONUNCIATIONS

Jing Zheng, Horacio Franco, Fuliang Weng, Ananth Sankar and Harry Bratt

Speech Technology and Research Laboratory
SRI International
http://www.speech.sri.com

ABSTRACT

Variations in rate of speech (ROS) produce changes in both spectral features and word pronunciations that affect ASR systems. To cope with these effects, we propose to use rate-specific phone models and pronunciations for ROS modeling at the word level. Words are given three types of pronunciations — fast, slow, and medium — consisting of rate-specific phone models, respectively. This approach allows us to model within-sentence rate variation. To better model coarticulation effects, we introduce the concept of zero-length phones, which enables short phones to be skipped without having to change their neighboring phones' contexts. A data-driven approach is used to prune the pronunciation dictionary derived from rules for phone reduction. We tested these approaches on the Hub 4 database and achieved a relative improvement of 2.0% over the baseline — an evaluation-quality version of SRI's DECIPHER™ continuous speech recognition system — for clean native speech in the 1996 development set.

1. INTRODUCTION

People have observed that speech rate significantly affects the performance of ASR systems [1-5]. Speech rate may influence the spectral features [2], and it may also influence the pronunciation [1]. In previous work [4], an input utterance was first classified as fast or slow by using a rate-of-speech (ROS) detector, and then fed to a rate-specific system tuned to fast or slow speech. However, this approach presumes that the speech rate within an utterance is uniform, which is often not the case in conversational speech. In following sections we present evidence that the speech rate usually varies within a sentence.

The coarticulation phenomenon, which is highly rate dependent, obviously influences word pronunciation. In [1], the pronunciation dictionary was modified for fast speech according to linguistic rules; however, no improvement was obtained from this modification. This suggests that simple rule-based-only pronunciation transformation may not be a solution.

In this work, our basic approach is to use rate-specific acoustic models to characterize speech with different rates. To capture within-utterance rate variation, we calculate ROS at the word level rather than the sentence level, and give each word rate-specific multiple pronunciations consisting of rate-specific phones. The corresponding models are trained from pre-classified and tagged training data. The classification is based on a ROS measure at the word level. To model the coarticulation phenomenon, we propose a zero-length-phone concept, which enables some short phones to be skipped during the search without changing neighboring phones' contexts. This idea requires no change to our 3-state HMM topology, but allows us to model coarticulatory effects that occur around the deletion, or near-deletion, of a phone. To obtain a good pronunciation dictionary, after using some rules to generate possible pronunciations, we used a data-driven method to collect actually occurring pronunciations.

The experiments were performed on the Hub 4 database containing 200 hours of broadcast news. A version of the SRI DECIPHER CSR gender-dependent Genone [6] system with 240,000 Gaussians, and a bigram language model with a 48,000 word vocabulary is taken as the baseline.

2. THE ROS MEASURE

In our approach, the ROS measure is used to classify the training materials as fast, slow, or medium. Both alignment-based and signal-based measures are often used to calculate ROS [5]. Signal-based measures can be calculated prior to recognition, and thus could possibly be used as a guide for model selection. However, they are usually not as reliable as alignment-based measures. In this work, since the ROS measure is used only in training, we mainly study an alignment-based measure, which is computed from the duration of each word/phone in the alignments. As we want to use rate-specific phones, we need to classify the training data in such a way that every rate-specific phone should have enough occurrences so that its model can be robustly trained. Hence, we use a kind of relative ROS measure $R_w(D)$, as follows:

$$R_w(D) = \sum_{d=D+1}^{\infty} p_w(d) = 1 - \sum_{d=0}^{D} p_w(d), \qquad (1)$$

where w is a given word, D is the duration of w, and $p_w(d)$ the probability of that type of word having duration d. $R_w(D)$ is the probability of w having a duration equal or longer than D. The measure $R_w(D)$ always falls within the range [0,1]. In practice, phone duration distributions are easy to estimate, while word duration distributions are hard because of the sparseness of the training data. To address this we assume that in a word the component phones' duration distributions are independent of

each other. Thus, a word's duration distribution equals the convolution of its component phones' distributions.

In our experiment, the duration distribution of each type of monophone is obtained from analyzing forced Viterbi alignments dumped by the baseline system. Then, the duration distributions of all words that appear in the training data are calculated, and their rates are computed according to Eq. (1). In terms of these rates, the 30% fastest words are labeled with a "fast" tag and the 30% slowest with a "slow" tag. We found that 88% of the sentences have at least one word with a fast tag and other word with a slow tag; this clearly indicates that the ROS within an utterance is usually not uniform.

Fig. 1 illustrates the duration distribution of different types of phones. It is apparent that the distributions are quite different; this supports our motivation for using a relative ROS measure rather than an absolute one. We can also notice that some short phones have a high probability of a duration of three frames. This suggests that the HMM topology with 3-states and no skips across states in our baseline DECIPHER system is probably limiting the observed minimum duration of certain phones. Obviously, this problem is more severe in fast speech than in slow speech.

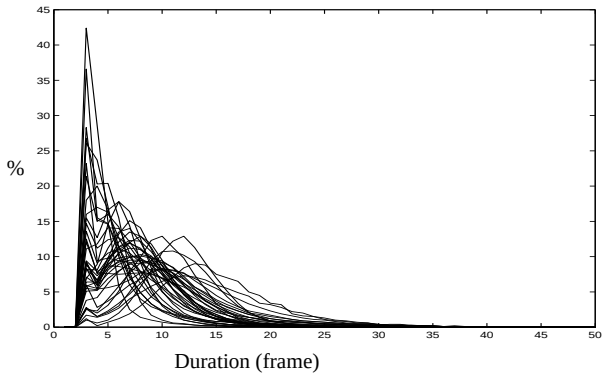

Figure 1: Duration distributions of different phone types

3. ROS-DEPENDENT MODELS

Since speech rate influences the spectral features, we want to use rate-specific phone models to describe speech with different rates. We studied the phone's relative ROS variation within words vs. within sentences using a measure similar to Eq. (1). In Fig. 2, we show a histogram of the variance of the ROS measure within words and within sentences for all Hub 4 training data. Fig. 2 indicates two things: (1) the word is a better unit than the sentence for ROS modeling because the rate variation within a word is significantly smaller than within a sentence; (2) inside the same word, different phones' rates are not totally independent (otherwise, intra-word rate variance should be equal to inter-word rate variance). Thus, we can impose word-level consistency by using word-level ROS. According to this analysis, we give a word three rate-specific pronunciations (assuming originally it had only one pronunciation), consisting of rate-specific phones. In the first step, we keep the original phone structure and clone it into three different versions. For example, the original pronunciation of "WORD" is /w er d/. We give it three pronunciations as /w_f er_f d_f/, /w_m er_m d_m/, and /w_s er_s d_s/, which consist of fast, medium and slow phones, respectively. At the same time, we clone the phonetic models, and create rate-specific models that are three times as large as the original models. The cloned models with the same origin share the same Genone [6].

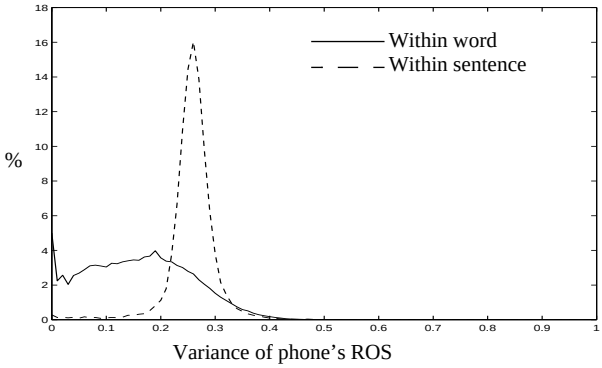

Figure 2: Rate variance distributions of a phone: within words vs. within sentences

3.1. Training

According to the ROS-labeled transcription, we modify the dumped Viterbi alignments and replace the original phones with rate-specific phones in terms of the corresponding word tag. Then, from these alignments, we train the rate-specific models using DECIPHER's Viterbi training tools. We do Bayesian smoothing between rate-specific models and the original models, which allows the parameters to be robustly estimated on a smaller set of rate-specific training data. Thus, from this training, we get a large set of phonetic models containing three versions for each triphone: fast, medium, and slow. Our language model uses a 48,000 vocabulary, in which many words never appear in the training data. For these words, we do not change the pronunciations, and use original rate-independent phones for the sake of robustness.

3.2. Experiments

We used the '96 DARPA Hub 4 development set (DEV) for evaluation. The data is split along the so-called F-conditions: prepared speech (F0), spontaneous speech (F1), low-fidelity speech (F2), speech with music background (F3), speech with noise (F4), non native speech (F5), and all other speech (FX). Since we did not integrate any technology to deal with music and noise, we focus only on the F0 and the F1 conditions. Table 1 compares the word error rate (WER) of the system with rate-dependent models and the baseline system with rate-independent models. The models have not been adapted to the testing data in any experiments described below. As we can see, rate-dependent model leads to a relative win of 1.7% over the rate-independent model, and as we expected, the improvement in F1 condition is larger than in F0 condition.

We also tested a signal-based ROS measure — *mrate* from ICSI [5]. In the experiment, a 1-second window is used to extract local *mrate* from the speech: the window advances one frame per step,

	F0 Male 2528 w.	F1 Male 4278 w.	F0 Female 1910 w.	F1 Female 1334 w.	Average 10050 w.
A	17.44	33.24	20.31	34.86	27.02
B	17.17	32.84	19.84	33.81	26.56

Table 1: Comparison of systems with: rate-independent models (A) and rate-dependent models (B) on 96 Hub 4 DEV data.

and thus we get the *mrate* for every frame. Then, still based on forced alignment, we compute the rate for each word by averaging the local *mrate* of all the frames the word covers. Using this rate measure, we replicated the same clustering and training steps, which resulted in another system. Table 2 compares the results of the two systems on the male data of the F0 and the F1 conditions, and suggests that our alignment-based measure is slightly better than the *mrate*-based measure in this case. (To save time, some of the minimum comparisons in different experiments were conducted over only the male data set.)

ROS Measure	F0	F1
Alignment-based	17.17	32.84
Mrate-based	17.64	32.87

Table 2: Comparison of systems trained with different ROS measures: alignment-based vs. *mrate*-based on 96 Hub 4 DEV male data.

4. PRONUNCIATION MODELING

Due to pronunciation reduction, some short phones in normal pronunciation can be deleted to fit fast speech better [1]. We have also observed the minimum duration problem of our HMM models in Fig.1, which suggests that removing some short phones from pronunciation may be a solution. In [1], some phone-deletion rules were applied to a pronunciation dictionary to generate new pronunciations that are supposed to be more suitable to fast speech. However, no improvements were obtained in the experiment. The authors addressed the problem that the models had not been retrained with the new pronunciation dictionary. We believe two other problems exist. First, most state-of-the-art ASR techniques use context-dependent phone models (e.g., triphones); simple deletion of a phone causes context changes in neighboring phones, and thus may result in some strange context-dependent phone models that have too little data for robust training; second, a rule-based method sometimes cannot fit training data very well, and some generated pronunciations do not actually happen and may increase inter-word confusability.

4.1. Zero-length Phone

We believe that in highly co-articulated pronunciation, though a phone may be too short to be represented as a normal phone, its influence on its neighboring phones' pronunciations should still be kept. Thus, we propose the zero-length-phone concept: allowing some phones to be skipped (having zero length) during the search, but at the same time, keeping the same context-dependent models for their neighboring phones as if they had not been skipped. For example, suppose a word's pronunciation is *a, b, c, d*. In the recognition system, the corresponding context-dependent model string is #[a]b, a[b]c, b[c]d, c[d]#. Now suppose phone *b* is too short to be preserved, the zero-length phone rule will result in the context-dependent model string #[a]b, b[c]d, c[d]#, where *b* is the zero-length phone that is actually skipped, while the common phone deletion scheme will lead to #[a]c, a[c]d, c[d]#. Obviously, in the former case, a clue of *b* is still kept, while in the latter, information about *b* is removed from the models. In addition, in the former case, no new context-dependent models will be required, while in the latter, #[a]c and a[c]d are introduced due to the change of contexts, and these models may be unlikely and have too little data to be trained robustly. For these reasons, we believe the zero-length phone concept is a better approach than common phone deletion.

4.2. Pronunciation Dictionary

The key problem for pronunciation modeling with zero-length phones is to decide which phones in a pronunciation should have zero length. Other approaches [1] allowed deletion of schwa. Based on preliminary analysis and recognition experiments we chose to focus on a subset of the consonants. We selected /d/, /dh/, /n/, /t/, /hh/, /v/, /y/, /r/, /g/, and /th/, whose mean durations were among the shortest, and allowed each to be a zero-length phone in any position in a pronunciation except the beginning or end of a word. Applying these rules to the original lexicon, we get a large pronunciation dictionary. Table 3 compares the performance of the system with the new dictionary and the baseline.

System	F0	F1
Baseline	17.44	33.24
Pronunciation dict. generated by rules	17.68	33.29

Table 3: WER comparison of the system with pronunciation dictionary generated by rules and the baseline system on 96 Hub 4 DEV male data.

It is clear that this rule-based method does not bring any improvement, but some degradation instead. We believe that this is due to the confusability brought by the drastically expanded pronunciation dictionary: applying these rules may produce a lot of wasteful pronunciations that will do harm to the recognition accuracy.

Since a lot of training data is available, a data-driven method [7] can be used to choose the useful pronunciations from the large lexicon generated by the rules. Using the recognizer with the large dictionary, we dump the forced Viterbi alignments for all training utterances, and then collect the actually occurring pronunciations for every word. For the sake of robustness, only those pronunciations that appear more than once are kept, and thus a new lexicon is obtained. For safety, we combine the resulting dictionary with the original, and obtain the final one. We build a recognition system using this dictionary and compare

it with the baseline, and this time we do get a tiny relative win of 0.4%, as shown in Table 4.

	F0 Male	F1 Male	F0 Female	F1 Female	Average
A	17.44	33.24	20.31	34.86	27.02
B	17.52	33.15	19.84	34.86	26.92

Table 4: Comparison of baseline (A) and data-driven pronunciation modeling method (B) on 96 Hub 4 DEV male data.

We also did an experiment to test the effectiveness of the zero-length-phone method versus common phone deletion. By simply changing the rules from allowing each of the 10 phones to be zero-length, to allowing them to be deleted in the conventional fashion, we replicated all the steps of the above experiment with the same kind of data-driven method. This procedure led to another recognition system, which was compared with the system based on the zero-length-phone approach described above. Results are shown in Table 5, which suggests that the zero-length-phone-based approach is better than the common phone-deletion-based approach.

Systems	F0	F1
PM with zero-length phone	17.52	33.15
PM with common phone deletion	18.04	34.32

Table 5: Comparison of pronunciation modeling (PM) method with zero-length phone and common phone deletion on 96 Hub 4 DEV male data.

	F0 Male	F1 Male	F0 Female	F1 Female	Average
A	17.44	33.24	20.31	34.86	27.02
B	17.25	32.42	19.79	34.56	26.49

Table 6: Comparison of baseline (A) and the system with rate-dependent models and rate-dependent pronunciations (B)

4.3. Rate-specific Pronunciation Modeling

The work we have described above was done on rate-independent models. As we know, pronunciation is relevant to speech rate, and coarticulation is more critical in fast speech than in slow speech; accordingly, rate-dependent pronunciation modeling seemed to be a promising approach. We apply the data-driven zero-length-phone-based pronunciation modeling method to our rate-dependent system described in Section 3. Thus, for different speech rates, the phonetic models are different, and the pronunciations are optimized differently. We do notice that the zero-length phones happen more frequently in "fast pronunciations" than in "slow pronunciations". The recognition results are listed in Table 6. As expected, the result is better than those of using rate-dependent models and zero-length-phone-based pronunciation modeling alone. The improvement over the baseline is 2.0% relative.

Though the proposed pronunciation modeling method does not bring a big improvement, we find the HMM likelihood score increase for the correct hypotheses with respect to the baseline system. This suggests that the new pronunciation dictionary matches the speech better. To obtain a larger gain, we need to study how to further reduce the inter-word confusability of the augmented dictionary, on the basis of the proposed data-driven method.

5. SUMMARY

We use rate-specific phone models and word pronunciations for rate-of-speech (ROS) modeling. By modeling at the word level, rather than sentence level, with ROS-specific models, this method accommodates within-sentence rate variation. We use an alignment-based rate measure to classify training data into fast, slow, and medium categories. A relative win of 1.7% is obtained by using rate-dependent models with respect to rate-independent models.

We propose the concept of zero-length phones to model phone reduction, which seemed to be better than the common phone-deletion scheme, and a data-driven method to prune the pronunciation dictionary generated by rules, which works better than a rule-based-only approach. The pronunciation modeling brings 0.4% relative gain based on rate-independent models. Combining rate-dependent models and the pronunciation modeling approach produces a relative win of 2.0%.

6. ACKNOWLEDGMENT

We gratefully acknowledge the support from SRI IRD funding, and Nelson Morgan at ICSI for making available the *mrate* code.

7. REFERENCES

[1]. M.A. Siegler and Richard M. Stern, "On the Effects of Speech Rate in Large Vocabulary Speech Recognition Systems," *ICASSP'95*, pp. 612-615, 1995

[2]. N. Mirghafori, E. Fosler and N. Morgan, "Fast Speakers in Large Vocabulary Continuous Speech Recognition: Analysis & Antidotes," *EUROSPEECH*'95, pp. 491-494, 1995

[3]. N. Morgan, E. Fosler and N. Mirghafori, "Speech Recognition Using On-line Estimation of Speaking Rate," *EUROSPEECH'97*, pp. 2079-2082, 1997

[4]. N. Mirghafori, E. Fosler and N. Morgan, "Towards Robustness to Fast Speech in ASR," *ICASSP'96*, pp. I335-338, 1996

[5]. N. Morgan and E. Fosler, "Combining Multiple Estimators of Speaking rate," *ICASSP'98*, pp. 729-732, 1998

[6]. V.V. Digalakis, P. Monaco and H. Murveit, "Genones, Generalized Mixture Tying in Continuous Hidden Markov Model-based Speech Recognizers," *IEEE TSAP*, vol 4. no 4. pp. 281-289, 1996

[7]. G. Tajchman, E. Fosler and D. Jurafsky, "Building Multiple Pronunciation Models for Novel Words Using Exploratory Computational Phonology," *EUROSPEECH'95*, 1995