

# Rate of Speech Modeling for Large Vocabulary Conversational Speech Recognition

*Jing Zheng, Horacio Franco, and Andreas Stolcke*

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 automatic speech recognition (ASR) systems. In order to deal with these ROS effects, we propose to use parallel, rate-specific, acoustic models: one for fast speech, the other for slow speech. Rate switching is permitted at word boundaries, to allow modeling within-sentence speech rate variation, which is common in conversational speech. Due to the parallel structure of rate-specific models and the maximum likelihood decoding method, we do not need a high quality ROS estimator before recognition, which is usually hard to achieve. In this paper, we evaluate our approach on a large-vocabulary conversational speech recognition (LVCSR) task over the telephone, with several minimal pair comparisons based on different baseline systems. Experiments show that on a development set for the 2000 Hub-5 evaluation, introducing word-level ROS-dependent models results in a 3.4% relative win over a baseline system without multiword pronunciation modeling, and a 1.8% relative win over a baseline system, that incorporates a 7.5% relative win from multiword pronunciation modeling.

## 1. INTRODUCTION

Rate of speech (ROS) is an important factor that affects the performance of a transcription system [1],[2]. Possible reasons are: some features commonly used in recognition systems are duration related and clearly influenced by speech rate, such as delta and delta delta features; some pronunciation phenomena such as coarticulation and reduction are also speech rate related. Thus, using rate-dependent acoustic models seems to be a promising way to improve robustness against speech rate variation.

In previous research work, rate-dependent acoustic models were often used at the sentence level. In the typical framework, an input utterance was first classified as fast or slow using a ROS estimator, and then fed to a rate-specific system that was tuned to fast or slow speech [2]. There are

two drawbacks for this framework, first, this approach presumes that the speech rate within an utterance is uniform, which is often not the case in conversational speech. In our earlier research work on broadcast news [3], we found that speech rate variation within sentences is common, and thus we proposed to use a more local rate-dependency for the acoustic models. Second, this framework is based on sequential classification, so errors on the first ROS classification will likely trigger errors in the recognition step. This paper propose a new framework of word-level rate-dependent acoustic modeling. Under this framework, each typical word is given two parallel rate-specific pronunciations: a fast-version pronunciation and a slow-version pronunciation, consisting of rate-specific phones respectively. The recognizer is allowed to select the fast or the slow pronunciation for each word automatically during search, based on the maximum likelihood criterion. This way, we can model the within-sentence speech rate variation, and avoid the requirement of pre-recognition ROS classification. To train the rate-specific phone models, we use a duration-based ROS measure to partition the training data into rate-specific categories. Due to the availability of training transcriptions, robust and accurate ROS estimation for training data can be achieved.

In Section 2 we will first introduce the ROS measure used for partitioning the training data. In Section 3 we will show the experimental result of rate-dependent acoustic modeling based on SRI's 98 evaluation system, and compare different training approaches. In Section 4 we will describe the work for the LVCSR 2000 (Hub 5) evaluation system, and specifically address the effect of multiwords in rate-depend acoustic modeling. Finally, in section 5, we will summarize our results.

## 2. ROS MEASURE

There are two typical methods to estimate ROS of an input utterance. One is based on phone durations, which are often obtained from phone-level segmentations by using forced alignments. When the utterance transcription is

known, this duration-based method was able to provide robust ROS estimation [2]; however, when the transcription is unknown, we can only use the hypothesis from an extra recognition run, whose quality is hard to guarantee. The second method is estimating ROS directly from the waveform or acoustic features of the input utterance [4]. To achieve robust ROS estimation, the computation is often based on a data window with sufficient length.

Under our proposed framework, in order to train the rate-specific models, we need to partition the training data into rate-specific categories at the word level, and we therefore need the ROS for each word to be estimated locally. The output of this process should give each word in the training transcription a rate class label. As our first step to ROS modeling, we decided to use only two ROS classes: fast or slow. Since we only need to compute ROS for the training data that have transcriptions, it is relatively straightforward to obtain the duration of each word and its component phones by computing forced Viterbi alignments, and then applying duration-based ROS estimation methods.

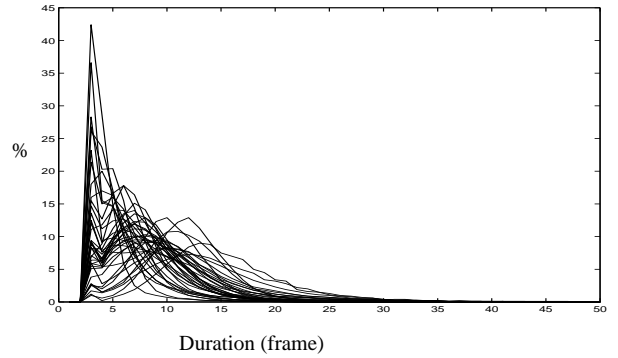
Absolute ROS measures, such as phones per second (PPS) and inverse mean duration (IMD) [2], were often used in previous work. However, we felt that these measures are not informative enough since they did not consider the fact that different type of phones have different duration distributions. Fig. 1 illustrates the duration distributions of 44 categories of monophones estimated from the 200 hours DARPA Hub 4 training corpus. As we can see, the duration distribution across different phone type differs substantially. If taking PPS or IMD as the ROS measure, words composed of short phones are easier to be treated as fast than those composed of long phones, even though they are not actually spoken faster than the normal rate. In our approach, we use a kind of relative ROS measure,  $R_w(D)$ , defined as follows:

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

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 longer than  $D$ . The measure  $R_w(D)$  always falls within the range  $[0,1]$ , and can be compared with each other between different word categories. However in practice,  $p_w(d)$  is hard to estimate directly due to the sparseness of the training data. To address this we assume that in a word the duration distribution of its component subword units, such as phones, are independent of each other. Thus, a word's duration distribution equals the convolution of its

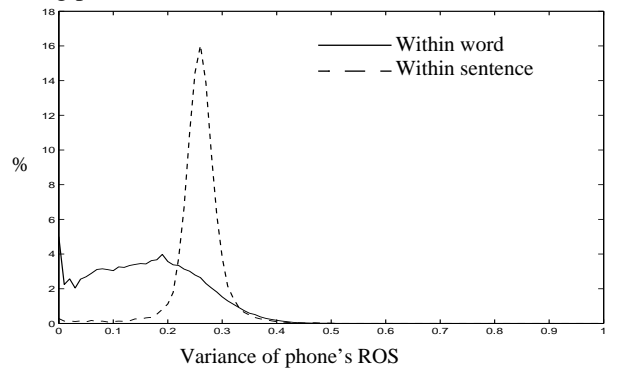
component subword units' distributions, which are easier to estimate from training data. In our recent research, we used triphones as the subword units for ROS estimation.

We used this measure to calculate the ROS for all the words in the training data of 200 hours broadcast news in the DARPA Hub 4 corpus, and found that 88% of the sentences have at least one word belonging to the 30% fastest words, and one word belonging to the 30% slowest words. This suggests that in conversational speech, speech rate is usually not uniform within a sentence.



**Figure 1:** Duration distributions of different phone types

In fact, the measure defined in Eq. (1) can also be applied to subword units, thus allowing us to calculate ROS of phones. Using this measure, we studied the phone's ROS variation within words vs. within sentences. Fig. 2 shows a histogram of the variance of the phone's ROS within words and within sentences for all Hub 4 training data, and it suggests two points: firstly, the word is a better unit than the sentence for ROS modeling because the mean rate variation within a word is significantly smaller than within a sentence; secondly, inside the same word, different phones' rates are not totally independent (otherwise, intra-word rate variance should be equal to inter-word rate variance). Hence, using word-level ROS and thus keeping word-level consistency seems to be a better approach than using phone-level ROS.



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

### 3. RATE-DEPENDENT ACOUSTIC MODELING

In our proposed method, each word is given parallel fast- and slow-version pronunciations in the recognition lexicon. Both fast- and slow-version pronunciations are cloned from the original rate-independent version, with the simple replacement of rate-independent phones by rate-specific phones. For example, the original rate-independent pronunciation of “WORD” is /w er d/. Then the fast-version pronunciation is /w<sub>f</sub> er<sub>f</sub> d<sub>f</sub>/, and the slow-version /w<sub>s</sub> er<sub>s</sub> d<sub>s</sub>/, which consist of fast, and slow phones, respectively. The recognizer is able to automatically find the best pronunciations that maximize the likelihood score during the search, and thus avoids the need of ROS estimation before recognition. In addition, the search algorithm is allowed to select pronunciations of different rate version across word boundaries, thus is able to cope with the problem of speech rate variation within sentence.

#### 3.1. Acoustic training

Our initial experiments are based on SRI’s 1998 Hub-5 evaluation system, which used continuous-density genonic Hidden Markov Models (HMM) [5]. The original evaluation system used a multi-pass recognition strategy [6], but for the sake of simplicity, we ran our experiments with only the first-pass recognizer, based on gender-dependent non-crossword genonic HMMs (1730 geneses with 64 Gaussians each for male, 1458 geneses for female), and a bigram grammar with a 33,275 word vocabulary. The recognition lexicon was derived from the CMU V0.4 lexicon with stress information stripped. The recognizer used a two-pass (forward pass and backward pass) Viterbi beam search algorithm, and in the first pass a lexical tree is used in the grammar backoff node to speed up search. Below we report results from the backward pass. The features used were 9 cepstral coefficients (C1-C8 plus C0) with their first and second order derivatives in 10ms time frames. The acoustic training corpus containing 121K male sentences and 149K female sentences came from (A) Macrophone telephone speech (B) 3094 conversation sides from the BBN-segmented Switchboard-1 training set (with some hand-corrections), and (C) 100 CallHome English training conversations.

We first calculated the ROS for all the words in the training corpus based on the above-mentioned measure, sorted these words accordingly, then split them into two categories: fast and slow. The ROS threshold for splitting is selected to achieve equal amounts of training data for the fast and the slow speech. The training transcriptions were labeled accordingly. We then prepared a special training lexicon: words with a fast label were given the fast-version

pronunciation, and words with a slow label the slow-version pronunciation. In this way, we were able to train the fast and slow models simultaneously.

We used DECIPHER genonic training tools to do standard MLE (Maximum Likelihood Estimation) training [5], and obtained rate-dependent models with 3233 geneses for male and 2501 geneses for female. The gene clustering for rate dependent models used the same information loss threshold as the training of rate-independent models.

We compared the rate-dependent acoustic model with the rate-independent acoustic model (baseline system) on a development data set, which is a subset of the 1998 Hub-5 evaluation data set, consisting of 1143 sentences from 20 speakers (9 male, 11 female). Table 1 shows the Word Error Rate (WER) for both models.

	male	female	all
rate-independent model	55.3	63.4	59.8
rate-dependent model from training	52.9	61.9	57.9

**Table 1:** WER comparison between the baseline system with rate-independent model and the system with rate-dependent model from training on the development data set

Rate-dependent modeling brings a relative WER reduction of 3.1%, which is statistically significant. To eliminate the possible effect of different numbers of parameters, we adjusted the information loss threshold for gene clustering to obtain another rate-independent model that had a similar number of parameters as the rate-dependent model. However, we did not observe any improvement from the increased number of parameters. This suggests the win is indeed from the introduction of rate dependency.

#### 3.2. Adaptation vs. standard training

In our previous work on the Broadcast News corpus (Hub 4) [3], instead of the training method described above, we trained the rate-dependent model based on a modified Bayesian adaptation scheme [7], by adapting the rate-independent model to rate-specific data to obtain rate-specific models. This was motivated by the small amount of available training data relative to the model size. In [3], we used a baseline system with a very large model comprising 256K Gaussians, and classified the training data into three categories: fast, slow, and medium. For this model size the training data was not sufficient to perform standard training. However, in the current task of Hub-5 telephone speech transcription we had significantly more training data, and we used a different strategy to partition the data into two classes instead of three, yielding more training data for each rate class. In addition, the optimal

models we started with were smaller. Thus, we were able to train the rate-dependent model robustly with standard training methods. For comparison we tested the Bayesian adaptation approach that we used in [3] on the current training set. Similar to [3], even though we used separate rate-specific models for each triphone, we did not create separate copies of the genones, but let the fast and slow models for a given triphone share the same genone. In this way, we used the same number of Gaussians for the rate-dependent model as for the rate-independent model.

Table 2 shows the results on the same development data set we used in the previous section. As we can see, this approach brings a relative win of 1.7% over the baseline, but not as good as the standard training scheme. This indicates that the difference between fast and slow speech in the acoustic space is significant, and that standard training might be better than the previous adaptation scheme to capture this difference. In fact, standard training optimizes the parameter tying for the rate-dependent model, reestimates the HMM transition probabilities, and performs multiple iterations of parameter reestimation, while the adaptation approach does not recompute genonic clustering, does not change the transition probabilities, and includes only one iteration of reestimation for the rate-dependent model on top of the rate-independent model. These differences might explain why the adaptation scheme did not achieve as much improvement as the standard training.

	male	female	all
rate-independent model	55.3	63.4	59.8
rate-dependent model from adaptation	54.0	62.6	58.8

**Table 2:** WER comparison between the baseline system with rate-independent model and the system with rate-dependent model from adaptation on the development set.

#### 4. EXPERIMENTS IN THE LVCSR 2000 EVALUATION SYSTEM

For the March 2000 NIST Hub-5 benchmark numerous improvements were made to SRI’s 1998 evaluation system [8], and the baseline system had been augmented substantially. Below we show some minimal pair experiments based on different baseline systems during the development process. The baseline system in Table 3 used a wider-band front end (with 13 cepstral coefficients instead of 9), and Vocal Tract Length (VTL) normalization [9] during training. As we can see, the win from introducing word-level rate dependency is 3.4% relative. This improvement is even larger than in our previous

results, over a baseline that was itself improved by almost 9% relative.

	male	female	all
WER of baseline system	50.6	57.9	54.6
WER of rate-dependent system	49.2	55.6	52.7

**Table 3:** Minimal pair comparison based on an improved baseline system using a wider front end and VTL normalization on the development set.

Another major addition to the evaluation system was the introduction of multiword pronunciation modeling. Here multiword refers to high frequency word bigrams and trigrams, such as “a lot of”, which are handled as a single word in the vocabulary. By using handcrafted phonetic pronunciations describing various kinds of pronunciation reduction phenomena for these multiwords, we achieved better modeling of crossword coarticulation. In SRI’s 2000 evaluation system, 1389 multiwords were introduced. Experiments showed that the multiword pronunciation modeling brought about a 7.5% relative win on top of the improved baseline system in Table 3.

We tried applying our rate-dependent modeling approach to the multiword-augmented baseline system by treating the multiwords as ordinary words. In this case, we obtained a smaller win of 1.4% relative, as shown in Table 4. (Compared to Table 3, a small part of the baseline WER reduction -- about 1.4% absolute -- of the comes from other improvements, such as variance normalization and automatically cleaned-up training transcriptions.)

	male	female	all
WER of baseline system	44.4	53.0	49.2
WER of rate-dependent system	43.1	52.8	48.5

**Table 4:** Minimal pair comparison based on a multiword-augmented baseline system on the development set.

The possible reasons for the reduced improvement may lie in the following aspects. First, each multiword is given multiple parallel pronunciations reflecting both full and reduced forms. This by itself models fast and slow speech variants to some extent. However, since this affects only the 1389 multiwords, there should still be room for improvement from rate-dependent modeling. Second, by treating multiwords as ordinary words, we can not model the rate variation occurring within the multiwords, and thus may influence the quality of rate-dependent acoustic model. Third, due to our current implementation, the introduction of multiwords made the search much more expensive than before; rate-dependent modeling on top of

the multiword dictionary made this problem even worse, and may have produced a loss in performance due to search pruning.

Based on the above analysis, we tested another scheme: instead of treating multiwords as ordinary words we trained them with multiword-specific phone units, that is, using separate phonetic models to describe the multiwords. Similar to the original approach, we trained three classes of phone models simultaneously: fast models for ordinary words, slow models for ordinary words, and a separate set of phone models trained only on the multiword data. This way we obtained a larger WER reduction of 1.8% (relative), shown in Table 5. (Note the baseline system in Table 5 was not trained on the cleaned up transcription, and thus is slightly worse than the system in Table 4.).

	male	female	all
WER of baseline system	44.7	53.5	49.6
WER of rate-dependent system	43.6	52.7	48.7

**Table 5:** Minimal pair comparison on the development set between the multiword-augmented baseline system and the rate-dependent system with multiword-specific phone models

We replicated the same experiment on the 2000 Hub-5 evaluation data set, which contains 4466 sentences from 80 speakers (29 male, 51 female), also obtaining a win of 1.8% relative (which is statistically significant for this data set).

	male	female	all
WER of baseline system	40.4	42.2	41.6
WER of rate-dependent system	40.0	41.3	40.9

**Table 6:** Minimal pair comparison on the evaluation set between the multiword-augmented baseline system and the rate-dependent system with multiword-specific phone models

## 5. CONCLUSIONS AND FUTURE WORK

We proposed a rate-dependent acoustic modeling scheme, which is able to model within-sentence speech rate variation, and does not rely on ROS estimation prior to recognition. Experiments show that this method results in a 3.4% (relative) word error rate reduction on a Hub-5 telephone speech transcription test set. When combined with multiword pronunciation modeling, our method led to a win of 1.8% (relative) on the same data set, and a statistically significant win of 1.8% relative on the LVCSR 2000 evaluation set.

Our current approach uses identical pronunciations but different phone units to model fast versus slow speech. We are currently investigating several alternative approaches, such as making both phones and pronunciations rate-specific, and a more general way to account for crossword pronunciation variation that does not require multiwords.

## 6. REFERENCES

- [1]. M.A. Siegler and Richard M. Stern, "On the Effects of Speech Rate in Large Vocabulary Speech Recognition Systems," Proc. *ICASSP'95*, pp. 612-615, 1995
- [2]. N. Mirghafori, E. Fosler and N. Morgan, "Towards Robustness to Fast Speech in ASR," Proc. *ICASSP'96*, pp. 1335-338, 1996
- [3]. J. Zheng, H. Franco, F. Weng, et. al., "Word-level Rate-of-Speech Modeling Using Rate-Specific Phones and Pronunciations," To appear in the Proc. of *ICASSP'2000*
- [4]. N. Morgan and E. Fosler, "Combining Multiple Estimators of Speaking rate," Proc. *ICASSP'98*, pp. 729-732, 1995
- [5]. 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
- [6]. H. Murveit, J. Butzberger, V. Digalakis and M. Weintraub, "Large-Vocabulary Dictation Using SRI's DECIPHER(TM) Speech Recognition System: Progressive-Search Techniques," Proc. *ICASSP'93*, Vol.2, pp. 319 -322, 1993
- [7]. V. V. Digalakis, and L. G. Neumeyer, "Speaker Adaptation Using Combined Transformation and Bayesian Methods," *IEEE TSAP*, vol 4. no 4. pp. 294-300, 1996
- [8]. A. Stolcke et al., "The SRI March 2000 Hub-5 Conversational Speech Transcription System," To appear in Proceedings of the May 2000 Speech Transcription Workshop.
- [9]. S. Wegmann, D. McAllaster, J. Orloff, and B. Peskin, "Speaker Normalization on Conversational Telephone Speech," Proc. *ICASSP'96*, Vol.1, pp.339-341, 1996