

RATE-DEPENDENT ACOUSTIC MODELING FOR LARGE VOCABULARY CONVERSATIONAL SPEECH RECOGNITION

Jing Zheng, Horacio Franco, and Andreas Stolcke

Speech Technology and Research Laboratory
SRI International
Menlo Park, CA 94025

ABSTRACT

Variations in rate of speech (ROS) produce changes in both spectral features and word pronunciations that affect automatic speech recognition (ASR) systems. 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 high-quality ROS estimation 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 1.9% absolute win over a baseline system without multiword pronunciation modeling, and a 0.7% absolute win over a baseline system that incorporates a 4.0% absolute 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 that some features commonly used in recognition systems are duration related and clearly influenced by speech rate, such as delta and delta delta features, and that 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]. This method has two

drawbacks. First, it 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 approach is based on sequential classification, so errors on the first ROS classification will most likely trigger errors in the recognition step. This paper proposes a new approach of word-level rate-dependent acoustic modeling. Under this approach, each typical word is given two parallel rate-specific pronunciations: a fast-version pronunciation and a slow-version pronunciation, each consisting of rate-specific phones. 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 first introduce the ROS measure used for partitioning the training data. In Section 3 we show the experimental results of rate-dependent acoustic modeling based on SRI's 1998 evaluation system, and compare different training approaches. In Section 4 we describe the work for the LVCSR 2000 (Hub 5) evaluation system, and specifically address the effect of multiwords in rate-dependent acoustic modeling. Finally, in Section 5, we summarize our results.

2. ROS MEASURE

Two methods are typically used 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 can provide robust ROS estimation [2]; however, when the transcription is unknown, we can only use the hypothesis from a prior recognition run, whose quality is hard to guarantee. The second method involves 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 approach, 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 types of phones have different duration distributions. Fig. 1 illustrates the duration distributions of 46 categories of monophones estimated from the training corpus. As we can see, the duration distribution across different phone types differs substantially. When taking PPS or IMD as the ROS measure, words composed of short phones are more easily 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 relative ROS measure, $R_w(D)$, defined as a percentile of a word’s ROS distribution:

$$R_w(D) = P_w(d > D) = 1 - \sum_{d=0}^D P_w(d), \quad (1)$$

where W is a given word, D is the duration of W , and $P_w(d)$ is 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 between different word categories. However in practice, $P_w(d)$ is hard to estimate directly due to the data sparseness problem. To address this we assume that in a word the duration distributions 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.

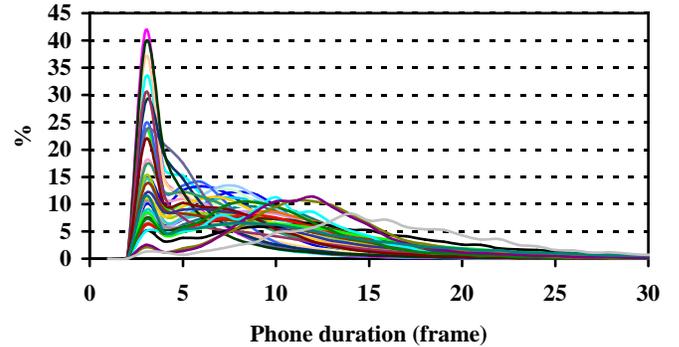
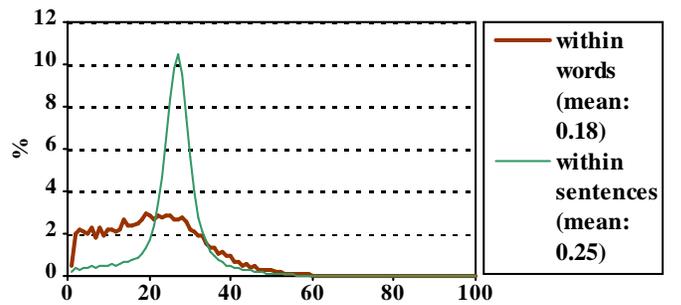


Figure 1: Duration distributions of different phone types

We used this measure to calculate the ROS for all the words in the training data, and found that 80% of sentences with five or more words have at least one word belonging to the fastest one third and one word belonging to the slowest one third of all the words. This suggests that in conversational speech, speech rate is usually not uniform within a sentence.

In fact, the measure defined in Eq. (1) can also be applied to subword units, thus allowing us to calculate the 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 standard deviation of the phone’s ROS within words and within sentences for all training data, suggesting that the word is a better unit than the sentence for ROS modeling, because the average phone-level ROS variation within a word is significantly smaller than within a sentence.



Standard deviation of phone-level ROS (100X)

Figure 2: Histogram of standard deviation of phone-level ROS: within words vs. within sentences

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 initialized 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/. Consequently the fast-version pronunciation is /w_f er_f d/ and the slow-version /w_s er_s d/, consisting of fast and slow phones, respectively. The recognizer automatically finds the best pronunciations that maximize the likelihood score during the search, and thus avoids the need for ROS estimation before recognition. In addition, the search algorithm is allowed to select pronunciations of different rates across word boundaries, thus coping with the problem of speech rate variation within a sentence.

3.1. Acoustic Training

Our initial experiments were based on SRI’s 1998 Hub-5 evaluation system, which uses continuous-density genonic hidden Markov models (HMMs) [5]. The original evaluation system used a multipass 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; in the first pass a lexical tree was 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 121,000 male sentences and 149,000 female sentences came from (A) Microphone telephone speech, (B) 3,094 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, and then split them into two categories: fast and slow. The ROS threshold for splitting was 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) gender-dependent training [5] and obtained rate-dependent models with 3233

geneses for male speech and 2501 geneses for female speech. 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. Note that all the results reported here are based on speaker-independent within-word triphone acoustic models and bigram language models, and are therefore not comparable with that of the full evaluation system.

	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 on the development data set

Rate-dependent modeling brings an absolute WER reduction of 1.9%, 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 number of parameters similar to that of the rate-dependent model in size. However, we did not observe any improvement from the increased number of parameters. This suggests the win is indeed due to 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 using 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 256,000 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 genome. 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. We see that this approach brings a win of 1.0% over the baseline, less than 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; whereas 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 2000 NIST HUB-5 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 enhanced 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 still 1.9%, over a baseline that was itself improved by 5.0%.

	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 pronunciations. A multiword is a high-frequency word bigram or trigram, such as “a lot of”, that is handled as a single unit 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 4.0% absolute win on top of the improved baseline system in Table 3, [8].

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 0.5% , as shown in Table 4. (Compared to Table 3, a small part of the baseline WER reduction -- about 1.3% absolute -- comes from other improvements, such as variance normalization and pronunciation probabilities.)

	male	female	all
WER of baseline system	44.3	53.3	49.3
WER of rate-dependent system	43.6	53.0	48.8

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

The possible reasons for the diminished effectiveness of ROS modeling 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 fail to model the rate variation occurring within the multiwords, and thus may influence the quality of the rate-dependent acoustic models. 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. With this approach, we improved the WER reduction to 0.7%, as shown in Table 5.

	male	female	all
WER of baseline system	44.3	53.3	49.3
WER of rate-dependent system	43.6	52.6	48.6

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

Finally, 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 0.7% absolute (which is statistically significant for this data set), as listed in Table 6.

	male	female	all
WER of baseline system	40.0	41.8	41.2
WER of rate-dependent system	39.7	41.0	40.5

Table 6: Minimal pair comparison on the 2000 NIST Hub-5 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 1.9% (absolute) 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 0.7% on the same data set, and a statistically significant win of 0.7% 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.

REFERENCES

- [1] M.A. Siegler and Richard M. Stern, "On the Effects of Speech Rate in Large Vocabulary Speech Recognition Systems," Proc. *ICASSP'95*, vol.1, pp. 612-615, 1995
- [2] N. Mirghafori, E. Fosler and N. Morgan, "Towards Robustness to Fast Speech in ASR," Proc. *ICASSP'96*, vol. 1, pp. 335-338, 1996
- [3] J. Zheng, H. Franco, F. Weng, A. Sankar and H. Bratt, "Word-level Rate-of-Speech Modeling Using Rate-Specific Phones and Pronunciations," Proc. *ICASSP'00*, vol 3, pp 1775-1778, 2000
- [4] N. Morgan and E. Fosler, "Combining Multiple Estimators of Speaking rate," Proc. *ICASSP'98*, vol 2, pp. 729-732, 1995
- [5] 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. 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," Proc. *NIST Speech Transcription Workshop*, college Park, MD, May 2000.
- [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