Effective Acoustic Modeling for Rate-of-Speech Variation in Large Vocabulary Conversational Speech Recognition

Jing Zheng *Horacio Franco* *Andreas Stolcke*

Speech Technology and Research Lab, SRI International
333 Ravenswood Ave., Menlo Park, CA 94025
{zj,hef,stolcke}@speech.sri.com

Abstract

We investigate several variants of speech-rate-dependent acoustic models for large-vocabulary conversational speech recognition, in the framework of combining rate-specific models in decoding to compensate for speech rate variation. We study two basic approaches to combining rate-specific models: one combines models at the pronunciation level and the other at the HMM state level. Furthermore, we investigate the influence of different numbers of rate-of-speech classes and different parameter tying schemes. Experiments on the Switchboard database, using SRI's DECIPHER recognition system, show that rate-dependent acoustic modeling resulted in a 2% relative word error rate reduction over a rate-independent baseline, and that the pronunciation-level constraint, Gaussian sharing between rate-specific models, and a well-chosen number of rate-of-speech classes are all important for best performance.

1. Introduction

Rate of speech (ROS) is known to be an important variable affecting performance of automatic speech recognition (ASR) systems. At the feature level, speech rate can affect especially the velocity and acceleration feature components. In speech production, the degree of coarticulation and phonetic reduction also correlates with speaking rate. Compensation for, or adaptation to, speech rate variation is therefore important for speech recognition, and has received extensive attention from researchers [1-6].

This paper focuses on modeling conversational speech, where within-sentence ROS variation is a common phenomenon [6]. In our prior research, we proposed a word-level rate-dependent acoustic modeling method, and obtained positive results on both the Broadcast News and Switchboard databases [5, 6]. The key idea behind this approach is to combine rate-specific acoustic models at the pronunciation level, and let the recognizer select the best matching models based on the maximum likelihood criterion during decoding. This approach addresses the issue of local speech rate variation and avoids the problems associated with automatic speech rate classification prior to recognition.

Motivated by some recent research in hidden pronunciation modeling [7, 8], we wanted to investigate an alternative approach of combining rate-specific acoustic models at the hidden Markov model (HMM) state level. In other words, we wanted to verify that the pronunciation-level constraint (requiring all phones in a word instance to belong to the same rate class) is important for our approach. Not implementing this constraint would simplify the modeling and reduce the recognition search space. We also wanted to examine the effect of different numbers of speech rate classes and different degrees of parameter sharing between classes. We used a more complex baseline system with a much lower word error rate than in our earlier work. The results are therefore indicative of the potential benefit of ROS modeling in the state of the art of large-vocabulary conversational speech recognition (LVCSR) technology.

The rest of the paper is organized as follows. Section 2 reviews the word-level model combination approach. Section 3 describes the state-level combination approach. Section 4 discusses experimental results, and Section 5 presents conclusions and describes plans for future work.

2. Word-level Model Combination

Before describing the model combination approach, we explain how the rate-specific models are trained. We classify all the words in an acoustic training transcription into the desired number of rate classes, using a local rate-of-speech measure. Each speech rate class is associated with a specific set of pronunciations and context-dependent phonetic HMMs. We then use SRI's genonic model training procedure [9] to obtain rate-specific models based on the labeled transcription. The rate-specific phonetic HMMs can either share or not share Gaussian densities between different rates, as determined by the training configuration.

The ROS measure for word-level speech rate classification is obtained from phone-level forced alignments of the training data. The ROS measure for a given word w with duration D_w (in frames), denoted $R_w(D_w)$, is defined as

$$R_w(D_w) \overset{\Delta}{=} 1 - P(d_w > D_w) = 1 - \sum_{d=1}^{D_w} P_w(d) \qquad (1)$$

where P_w (d) is the probability of an instance of word w having a duration of d frames. $P_w(d)$ can be estimated from the duration probability distribution functions of its constituent triphones, as described in Equation (2):

$$P_w(d) \approx \sum\sum\cdots\sum_{d_1+d_2+\cdots d_n=d} \prod_{i=1}^{n} P_{ph\_i}(d_i) \qquad (2)$$

where n is the number of triphones of word w; d_i is the duration of the i-th triphone; and *ph_i* is the name of the i-th triphone. $P_{ph\_i}(d_i)$ is the probability of triphone *ph_i* having duration d_i,, which can be directly estimated from statistics collected on the forced alignments database. Detailed information regarding this measure can be found in [6]. We sort all the words in the training transcriptions based on the speech rate computed as in Equation (1), divide them into a specified number of classes with equal numbers of words in each class, and label them with rate markers for acoustic training.

During recognition we combine all rate-specific models at the pronunciation level, which can be formulated as follows:

$$\begin{aligned} & p(X \mid w, M) = \sum_r P(r \mid w)\, p(X \mid w, M_r) \\ & \approx \max_r P(r \mid w)\, p(X \mid w, M_r) \\ & = \max_r P(r \mid w) \sum_v P(v \mid r, w)\, p(X \mid w, v, M_r) \\ & \approx \max_{r,v} P(v, r \mid w)\, p(X \mid w, v, M_r) \end{aligned} \quad (3)$$

where M represents the whole parameter set, w a given word, X the input speech feature vectors associated with w, r an arbitrary speech rate class, M_r the rate-specific acoustic model, and v the pronunciation variant. Equation (3) means that we use the best speech rate class pronunciation variants to compute likelihood scores; this is commonly known as the Viterbi approximation. In this way, a rate-dependent acoustic model can be easily integrated into a Viterbi search algorithm, which automatically selects the best rate class and pronunciation variant. We scaled the probabilities $P(v,\ r \mid w)$ such that $\max P(v,\ r \mid w) = 1.0$ in order to avoid penalizing words with multiple pronunciations.

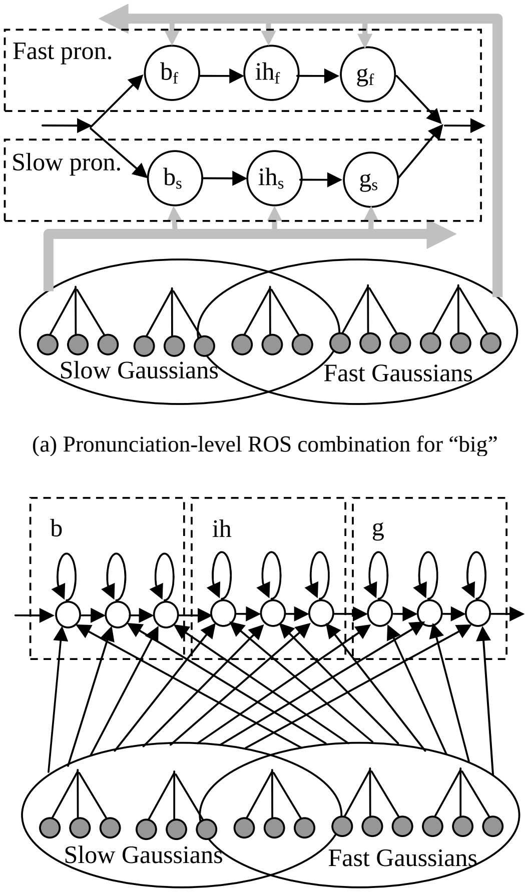
3. State-level Model Combination

Recent research indicates that pronunciation variation can be modeled implicitly without having to use multiple pronunciations. Saraclar et al. [8] found that state-level pronunciation modeling via sharing Gaussian densities across phones gave more improvement in accuracy than traditional phone-level pronunciation modeling. Hain [7] showed that with a suitable procedure for coalescing pronunciation variants, using a single-pronunciation dictionary can achieve performance that is similar to, or better than, that achieved with multiple pronunciations.

In light of these results, we also wanted to verify if the pronunciation-level constraint in our method is necessary. We developed an alternative approach which combines rate-specific models at the state level instead of at the word level. The rate-specific acoustic models are trained in the same manner as described in Section 2, but during recognition we use the original rate-independent pronunciations and phonetic HMMs. The observation probability density functions (pdfs) of the HMMs take the form of a mixture of pdfs corresponding to each speech rate class, which can be formulated as follows:

$$p(X \mid s, M) = \sum_r P(r \mid s)\, p(X \mid s, M_r) \quad (4)$$

Here M represents the whole parameter set, s a given HMM state of some phonetic model, X an input speech feature vector associated with s, r an arbitrary speech rate class , M_r the rate-specific acoustic model, and $p(X \mid s, M_r\ )$ the rate-specific pdf, taking the standard form of a Gaussian mixture model. Note that because we use rate-specific phonetic HMMs in training, we also have rate-specific HMM states. Equation (4) can be viewed as a special tying scheme that allows a rate-independent phonetic HMM state to be associated with Gaussian densities of all the corresponding rate-specific states with proper mixture weights. We call this tying scheme "super mixture". The probabilities $P(r|s)$ were initialized with uniform distributions, and then reestimated using an expectation maximization (EM) algorithm.



(a) Pronunciation-level ROS combination for "big"

(b) State-level ROS combination for "big"

Fig.1 Illustration of the two rate-specific model combination approaches: word-level vs. state-level.

In the standard tying scheme, one state is associated with one Gaussian mixture density function, while the rationale behind super-mixture tying is to better model speaking rate consistency, which is not appropriate to model at the frame level. By training Gaussians with rate-labeled transcriptions and rate-specific phonetic models, we position Gaussians suitably for speech at different rates. During recognition, although the pronunciation-level constraint is dropped, the Gaussians with super-mixture tying should still preserve some ability to model rate consistency. This is similar in spirit to the approach of hidden pronunciation modeling. By comparing state-level combination and word-level combination, we can ascertain the need for a pronunciation-level constraint that enforces the same rate throughout the word.

Figure 1 illustrates the two combination approaches, using the word "big" as an example. In this case two classes of speech rate are defined, fast and slow. Each class corresponds to 50% of the training data. The rate-independent pronunciation for "big" is /b ih g/. For the pronunciation-level combination approach, we duplicate the pronunciation into a "fast" version /b_f ih_f g_f/ and a "slow" version /b_s ih_s g_s/. They are associated with the "fast" and "slow" Gaussian pools, respectively. The fast and the slow Gaussian pools are trained using rate-labeled transcriptions and rate-specific pronunciations, and can either have some overlap, or be disjoint, as determined in the training procedure. Both schemes are evaluated. With the state-level combination, the pronunciation of "big" stays unchanged, though the HMM states will be associated with Gaussians from both fast and slow Gaussian pools, as expressed by Equation (4).

4. Experiments

Experiments were performed on a subset of the DARPA 2002 Rich Transcription (RT-02) evaluation database, consisting of around two hours of conversational telephone speech sampled at 8 kHz. The baseline system is a subsystem of SRI's full CTS evaluation system, which rescores lattices with maximum likelihood linear regression (MLLR) adapted cross-word-triphone acoustic models created with speaker-adaptive training (SAT). The acoustic models were trained on about 418 hours of data using gender-dependent speaker-adaptive maximum likelihood training, which resulted in 195,000-Gaussian male models and 175,000-Gaussian female models. Acoustic features included 13-dimensional Mel frequency cepstral coefficients (MFCC) with the first-, second-, and third-order derivatives, and 10 dimensions of voicing features [10], with vocal tract length normalization and mean and variance normalization. This 62-component feature vector was then projected onto 39 dimensions using a heteroscedastic linear discriminant analysis (HLDA) transform. Lattices generated from earlier decoding steps were expanded with trigram Super-ARV multiword language models for rescoring [11]. A 9-transform unsupervised MLLR adaptation was applied using hypotheses generated from the system's previous decoding steps.

We first investigated which of the two combination approaches is more effective for modeling rate dependency. In this experiment, we defined two speech rate classes, each containing 50% of the word tokens in the training set transcription. To make the comparison fair, we deliberately made the total parameter size of the rate-dependent model similar to that of the baseline rate-independent model. We allowed the two sets of rate-specific models to share Gaussians. In the case of state-level combination, we tested three conditions: initializing $P(r|s)$ in Equation (4) with uniform probabilities, in this case 0.5 for each speech rate class; reestimating $P(r|s)$ with one EM iteration; and reestimating both $P(r|s)$ and Gaussian parameters with one EM iteration. Note that the EM iteration used transcriptions without ROS labels. For comparison, in the case of word-level combination, we also ran an EM iteration using the combined pronunciations and transcriptions without ROS labels to jointly optimize parameters. Table 1 shows the results.

In Table 1 we see that both word-level and state-level combination approaches brought some WER reduction over the rate-independent baseline model; however, word-level combination is more effective, leading to about 2% relative WER reduction. This suggests that the pronunciation-level constraint is important in our modeling approach. While the state-level approach allows the models to better fit rate variation within words, the lack of a word-level constraint may well increase the confusability between words.

Table 1: Comparing word error rate (WER) of rate-independent model, rate-dependent model with word-level combination, and rate-dependent model with state-level combination. Two rate classes were used with Gaussian sharing.

Model training/combination		WER
Baseline, Rate-independent		25.5
Word-level Combination	ROS-labeled training	25.0
	+ 1 iteration Joint EM	24.9
State-level Combination	ROS-labeled training, uniform $P(r|s)$	25.4
	Reestimated $P(r|s)$	25.4
	Reestimated $P(r|s)$ and Gaussians	25.3

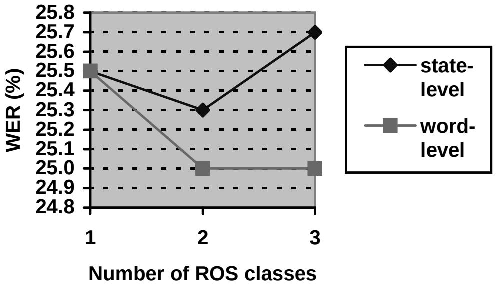


Figure 2: WER as a function of number of ROS classes

The results also show little difference between using uniform $P(r|s)$ and reestimated $P(r|s)$, probably because we used equal amounts of data for training each class of rate-specific models, and the dynamic range of $P(r|s)$ is very small compared to scores generated from evaluating Gaussian density functions. One extra EM iteration reestimating Gaussian parameters from transcriptions without ROS labels gave similar improvements for both word-level and state-level combination.

The second experiment explores the effect of the number of ROS classes on model accuracy, for both model combination schemes. Intuitively, using a larger number of ROS classes could improve a model's precision, but could also impair robustness, since less training data would be available for each rate class. The ideal number of ROS classes could be a function of the available amount of training data. In addition, varying the number of ROS classes could result in different interactions with different model combination approaches. In the word-level combination approach, a large number of ROS classes will significantly increase the search burden as the number of pronunciations per word increases. This is not the case for the state-level combination.

We tested the cases of one (rate-independent), two, and three ROS classes, with both combination approaches. Again, for fair comparison, in all cases the total parameter sizes were similar and Gaussian sharing between rates was allowed. Figure 2 shows the word error rate as a function of the number of ROS classes for the two combination approaches.

Figure 2 shows that using three classes of ROS instead of two did not give further improvement in the case of word-level combination, but led to a substantial degradation in the case of state-level combination. This shows that selecting an

Table 2: Comparing word error rate with and without cross-rate Gaussian sharing. Two rate classes were used.

Tying/Model Combination		WER
Rate-independent model		25.5
Share Gaussians across rates	Word-level combination	25.0
	State-level combination	25.4
Not share Gaussians across rate	Word-level combination	25.2
	State-level combination	25.5

appropriate number of ROS classes is important, and the word-level combination seems more robust to different choices of ROS classes than the state-level combination.

Finally, we wanted to know if it is important to allow different rate-specific models to share Gaussians, which had been the default setting in all of our previous experiments. In the experiment, we used two rate classes, and trained rate-specific models with and without cross-rate Gaussian sharing. The results are shown in Table 2.

From Table 2 it is easy to see that allowing cross-rate Gaussian sharing yields lower word error rates, in both the word-level combination and the state-level combination approaches.

5. Conclusions and Future Work

We have compared several different variations in rate-dependent acoustic modeling. From the experimental results, we can draw the following conclusions: First, training acoustic models conditioned on speaking rate helps reduce the word error rate of our ASR system at a level of 2% relative. Second word-level combination is a better way of utilizing rate-dependent models than state-level combination; in other words, it is important to constrain phones within the same word to have a consistent rate. Third, allowing models for different rate classes to share Gaussians is better than keeping them disjoint.

There is potential for improvements in our approach. First, the experiments showed us that appropriate parameter tying is important. In our standard Genonic training procedure, above the Gaussian level, parameters such as pronunciations, phonetic HMMs, and mixture weights are disjoint for different rates. We believe that by using decision tree clustering [12], we can tie parameters in a more seamless way, achieving a gradual transition from rate-specific models to rate-independent models, which could help improve modeling robustness as well as precision. Second, in the current approach, using too many ROS classes entails more search during recognition. Discriminative criteria [13] could reduce the number of pronunciations without a loss in recognition accuracy, and could be combined with discriminative acoustic training. Third, we have so far considered only within-word speaking rate consistency. Cross-word modeling of speech rate variation could be accomplished using language model techniques [14].

6. Acknowledgment

This work was funded by DARPA under contract No. MDA972-02-C-0038. Distribution is unlimited.

7. References

[1] Fosler-Lussier, E., and Morgan, N., "Effects of speaking rate and word frequency on pronunciations in conversational speech," *Speech Communication*, 29(2-4): 137-158, 1998.

[2] Mirghafori, N., Fosler E., and Morgan N., "Towards robustness to fast speech in ASR," *Proc. IEEE Int'l Conf. Acoust. Speech Signal Proc.*, Vol. 1, pp. 335-338, Atlanta, 1996.

[3] Morgan, N., and Fosler, E., "Combining multiple estimators of speaking rate," *Proc. IEEE Int'l Conf. Acoust. Speech Signal Proc.*, Vol. 2, pp. 729-732, Seattle, 1998.

[4] Siegler, M.A., and Stern, R.M., "On the effects of speech rate in large vocabulary speech recognition systems," *Proc. IEEE Int'l. Conf. Acoust. Speech Signal Proc.*, Vol. 1, pp. 612-615, Detroit, 1995.

[5] Zheng, J., Franco, H., Weng, F., Sankar, A., and Bratt, H., "Word-level rate-of-speech modeling using rate-specific phones and pronunciations," *Proc. IEEE Int'l Conf. Acoust. Speech Signal Proc.*, Vol. 3, pp. 1775-1778, Istanbul, 2000.

[6] Zheng, J., Franco, H., and Stolcke A., "Modeling word-level rate-of-speech variation in large vocabulary conversational speech recognition," *Speech Communication*, 41(2-3):273-286, 2003.

[7] Hain, T., "Implicit pronunciation modeling in ASR," *Proc. Pronunciation Modeling and Lexicon Adaptation for Spoken Language Technology*, pp. 129-134, Estes Park, 2002.

[8] Saraclar, M., Nock, H., and Khundanpur, S., "Pronunciation modeling by sharing Gaussian densities across phonetic models," *Computer Speech and Language*, Vol. 14, pp. 137-160, 2000.

[9] Digalakis, V., Monaco, P., and Murveit, H., "Genones, generalized mixture tying in continuous hidden Markov model-based speech recognizers," *IEEE Trans. Acoust. Speech Signal Proc.*, 4(4):281-289, 1996.

[10] Graciarena, M., Franco, H., Zheng, J., Vergyri, D. and Stolcke, A., "Voicing feature integration in SRI's DECIPHER LVCSR system," to appear in *Proc. IEEE Int'l Conf. Acoust. Speech Signal Proc.* Montreal, 2004.

[11] Wang, W., Stolcke, A., and Harper M., "The use of a linguistically motivated language model in conversational speech recognition," to appear in *Proc. IEEE Int'l Conf. Acousti. Speech Signal Proc*. Montreal, 2004.

[12] Paul, D., "Extensions to phone-state decision-tree clustering: Single tree and tagged clustering," *Proc. IEEE Int'l Conf. Acoust. Speech Signal Proc.*, Vol. 2, pp. 1487-1490, Munich, 1997.

[13] Schramm, H., and Beyerlein, P., "Discriminative optimization of the lexical model," *Proc. Pronunciation Modeling and Lexicon Adaptation for Spoken Language Technology*, pp. 105-110, Estes Park, 2002.

[14] Kessens, J.M., Wester, M., and Strik, H., "Improving the performance of a Dutch CSR by modeling within-word and cross-word pronunciation variation," *Speech Communication*, 29(2-4): 193-207, 1999.