Modeling Word-Level Rate-of-Speech Variation in Large Vocabulary Conversational Speech Recognition

Jing Zheng, Horacio Franco, Andreas Stolcke

Speech Technology and Research Laboratory

SRI International

Correspondence Address:

Jing Zheng
Speech Technology and Research Laboratory, SRI International
333 Ravenswood Avenue, Menlo Park, CA 94025, USA

Email: zj@speech.sri.com
Tel: (650) 859-6129
Fax: (650) 859-5984

Number of Page: 31 (including tables and figures)

Number of Tables: 3

Number of Figures: 3

Key Words: Rate-of-speech modeling; Large vocabulary conversational speech recognition; Pronunciation modeling

Modeling Word-Level Rate-of-Speech Variation in Large Vocabulary Conversational Speech Recognition

Jing Zheng, Horacio Franco, Andreas Stolcke

Speech Technology and Research Laboratory

SRI International

Received 2 February 2001; revised November 2001 and April 2002

ABSTRACT

Variations in rate of speech (ROS) produce variations in both spectral features and word pronunciations that affect automatic speech recognition systems. To deal with these ROS effects, we propose to use a set of parallel rate-specific acoustic and pronunciation models. Rate switching is permitted at word boundaries, to allow within-sentence speech rate variation, which is common in conversational speech. Because of the parallel structure of rate-specific models and the maximum likelihood decoding method, our approach does not require ROS estimation before recognition, which is hard to achieve. We evaluate our models on a large-vocabulary conversational speech recognition task over the telephone. Experiments on the NIST 2000 Hub-5 development set show that word-level ROS-dependent modeling results in a 2.2% absolute reduction in word error rate over a rate-independent baseline system. Relative to an enhanced baseline system that models crossword phonetic elision and reduction in a multiword dictionary, rate-dependent models achieve an absolute improvement of 1.5%. Furthermore, we introduce a novel method to modeling reduced pronunciations that are common in fast speech based on the approach of skipping short phones in the pronunciation models while preserving the phonetic context for the adjacent phones. This method is shown to also produce a small additional improvement on top of ROS-dependent acoustic modeling.

ZUSAMMENFASSUNG

Schwankungen in der Sprechgeschwindigkeit (“rate of speech”, ROS) beeinflussen sowohl die spekralen Eigenschaften als auch die Aussprache von Wörtern und betreffen somit die automatische Spracherkennung. Um diesen Effekten Rechnung zu tragen, verwenden wir mehrere parallele, ROS-spezifische akustische und Ausprachemodellen im Erkenner. Dabei sind ROS-Wechsel an Wortgrenzen erlaubt, so dass Anpassungen an ROS-Änderungen innerhalb eines Satzes möglich sind. Aufgrund der parallelen Struktur der ROS-spezifischen Modelle und der Verwendung der Maximum-Likelihood-Methode ist eine Bestimmung der ROS vor der Spracherkennung nicht notwendig, was typischerweise ein schwieriges Problem darstellt. Wir testen unsere Modelle in der Erkennung von Telefongesprächen. Experimente mit dem NIST 2000 Hub-5-Korpus ergaben eine absolute Verringerung der Wortfehlerrate von 2.2% bei Benutzung von ROS-abhängigen akustischen Modellen verglichen mit einem ROS-unabhängigen Baseline-System. Gegenüber einem verbesserten Baseline-System, in dem phonetische Elidierungen und Reduktionen an Wortgrenzen mittels Multiwörtern erfasst sind, ergibt ein ROS-abhängiges System eine absolute Verbesserung von 1.5%. Ausserdem stellen wir eine neue Methode zur Modelliering von reduzierten Aussprachevarianten, die oft bei schnellem Sprechen auftreten, vor. Dieses Verfahren erlaubt das Überspringen von kurzen Segmenten im Aussprachemodel, wobei jedoch der phonetische Kontext von Nachbarsegmenten erhalten wird. Diese Methode ergibt eine geringfügige zusätzliche Verbesserung der ROS-abhängigen akustischen Modelle.

Résumé

Les variations de vitesse d'élocution (ROS) affectent les indices spectraux du signal vocal et la prononciation; les systèmes de reconnaissance automatique de la parole y sont donc exposés. Afin de combattre ces effets, nous proposons d'utiliser en parallèle deux groupes de modèles acoustiques et de prononciation, adaptés en fonction de la vitesse d'élocution. Le choix entre ces deux groupes peut basculer à la frontière des mots afin de rendre compte en cours d'énoncé des variations de cette vitesse, courantes en parole conversationnelle. Grâce au parallélisme des deux groupes de modèles et à la méthode de décodage basée sur le maximum de vraisemblance, notre approche ne demande pas l'estimation de la vitesse d'élocution avant décision de reconnaissance, ce qui serait difficile à réaliser. Nous évaluons nos modèles sur une tâche de reconnaissance automatique de la parole téléphonique grand vocabulaire. Les expériences sur une configuration de développement NIST 2000 Hub-5s montrent que notre modélisation obtient 2,2% d'amélioration du taux de reconnaissance de mots comparé à un système de base ne comportant pas de traitement de la dépendance à la vitesse d'élocution. Par rapport à un système de base amélioré où la coarticulation et les élisions sont modélisées dans un dictionnaire de multi-mots, notre modélisation dépendante de la vitesse d'élocution obtient 1,5% d'amélioration.

Nous avons de plus introduit une nouvelle modélisation des réductions phonétiques, fréquentes dans la parole à débit rapide, où les phones courts peuvent être omis en tant que segment mais préservés en tant que contexte phonétique pour les phones adjacents. Cette approche a également permis une légère amélioration s'ajoutant à celle qu'obtient la prise en compte des variations de vitesse d'élocution.

1. INTRODUCTION

Rate of speech (ROS)[1] has been observed as an important factor that affects the performance of a transcription system; speaking either too fast or too slow would lead to higher word error rate (WER) (Siegler and Stern, 1995; Mirghafori et al., 1996) for a number of possible reasons. First, ROS is related to the degree of acoustic realization; changes in ROS would result in variation in both acoustic observations and underlying pronunciation baseforms. Furthermore, some features commonly used in recognition systems are duration related and clearly influenced by speech rate, such as delta and delta-delta features. For these reasons, much prior work employs rate-dependent acoustic models to reduce model mismatch and improve robustness against speech rate variation.

However, previous research addressed rate-of-speech effects mostly at the sentence or higher levels. In Mirghafori et al. (1996), an input utterance was first classified as fast or slow, using a ROS estimator, and then fed to a rate-specific system tuned to fast or slow speech. An obvious problem with this method is that errors in the ROS classification are likely to trigger errors in the recognition step because of model mismatch. In Richardson et al. (1999), utterances were normalized based on a ROS measure, by performing cepstral feature interpolation on the time axis. Both of the above approaches presume 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 speech (Zheng et al., 2000), we found that speech rate variation within sentences is common. However, local ROS is often hard to estimate robustly. Richardson et al. (1999) observed that although phone-level ROS normalization gave considerably larger improvement than sentence-level normalization when the correct phone sequence was known, it failed to yield any improvement when the correct phone sequence was unknown. This result indicates the potential benefit from modeling ROS at a

[1] We use the terms "rate of speech", "speaking rate", and "speech rate" interchangeably in this paper.

more local level, but also suggests that in order to realize the benefit, the problem of robust estimation of local ROS must be solved first.

We will address the local ROS estimation problem by using parallel rate-dependent acoustic and pronunciation models at the word level. Each word is given two groups of rate-specific pronunciations: one group of "fast" pronunciations and one group of "slow" pronunciations, each being implemented by rate-specific phone models.[2] The recognizer is allowed to select the fast or the slow pronunciation for each word automatically during search, based on the maximum likelihood criterion. In this way, we account for within-sentence speech rate variation, and avoid the requirement of prerecognition 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. Because of the availability of transcriptions in training, robust and accurate ROS estimation is not an issue in our approach. In an experiment with a multiword-augmented dictionary, we verified the importance of modeling ROS at the word level instead of at a more global level, especially the multiword level.

As observed by Siegler and Stern (1995), fast speech frequently produces changes in word pronunciation as well as in phone articulation. To address this, we explore a new method for modeling rate-dependent pronunciation variation. Based on the concept of a zero-length phone (Zheng et al., 2000), we enable models of some short phones to be skipped in the search without changing the phonetic contexts of their neighboring phones; thus, we are able to model the coarticulatory effects of those short phones. A data-driven algorithm is used to generate the rate-dependent pronunciation dictionary with zero-length phones automatically from alignment data. The method effectively allows words to have different pronunciations (or pronunciation probabilities) for different ROS.

The remainder of the paper is organized as follows. Section 2 introduces the ROS measure used for partitioning the training data. Section 3 reports experimental results with rate-

[2] In principle, we could group the words in any number of clusters based on a certain ROS measure; a typical choice would be a trichotomy of slow, normal, and fast speech. Because of the limited amount of available training data we chose to use only two clusters, which we refer to as "slow" and "fast".

dependent acoustic modeling and compares different training approaches. Section 4 introduces rate-dependent pronunciation modeling and reports results with rate dependency in both acoustic and pronunciation models. Section 5 addresses issues with ROS modeling in a multiword-augmented recognition system. Section 6 summarizes the work presented.

2. RATE-OF-SPEECH MEASURE

Two types of methods are typically used to estimate the ROS of an input utterance. The first is based on phone durations, which are usually obtained from phone-level segmentations via forced Viterbi alignments. When the utterance transcription is known, this duration-based method can provide robust ROS estimation (Mirghafori et al., 1996); 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. Morgan and Fosler (1998) developed a signal-processing-based measure, known as *mrate*, to estimate syllable rate for rapid speech detection. To achieve robustness, the computation must use a data window of sufficient length (1-2 seconds), which is generally too long for estimating local ROS. Tuerk and Young (1999) proposed another measure based on the Euclidean distance between successive feature vectors for modeling speaking rate, and showed some discriminative power in classifying fast and slow phones. However, they did not report experimental results using this measure in an automatic speech recognition (ASR) system.

Under our proposed approach, training the rate-specific models requires partitioning the training data into rate-specific categories at the word level; 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. We decided to use only two ROS classes, “fast” and “slow”, for several reasons. First, increasing the number of ROS classes would reduce the amount of training data in each class, which is not desirable for a large vocabulary task. Second, in our method, search complexity increases rapidly with the number of ROS classes, as the number of pronunciations is proportional to the number of classes.

[Figure 1] ⇦

Because we need to compute ROS only for the training data for which transcriptions are available, 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) (Mirghafori et al., 1996), were used in previous work. However, we felt that these measures are not suitable for our purposes since they do not consider the fact that different phone types have different duration distributions. Figure 1 illustrates the duration distributions of five typical phonemes, /d/, /p/, /ch/, /ih/ and /ay/, [3] as estimated from the training corpus. Clearly the duration distributions for different phone types differ substantially. Based on PPS or IMD as the ROS measure, words composed of short phones would seem inherently "faster" than those composed of longer phones, even when spoken at the same speaking rate. Therefore, we use a relative ROS measure, $R_W(D)$, defined as 1 minus the distribution function of the word duration considered as a random variable:

$$R_W(D) = P_W(d > D) = 1 - \sum_{d=0}^{D} P_W(d) \tag{1}$$

where W is a given word, D is the duration of W in frame units (the step size of the signal analysis window, 10 ms in our system), 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.

It is interesting to note that the ROS obtained from equation (1) has a close-to-uniform distribution since $R_W(.)$ can be viewed as a histogram-equalization transform (Gonzalez and Woods, 1992) mapping the word duration D to the range of [0,1] with an equalized histogram. However, in practice, $P_W(d)$ is hard to estimate directly because of data sparseness. To address this problem we assume that, within a word, the duration distributions of its

[3] We use OGIbet for phone labeling throughout the paper.

component subword units, such as phones, are independent of each other. Thus, the duration probability of a word equals the convolution of its component subword unit probabilities, which are easier to estimate reliably from training data. This can be formulated as

$$P_W(D) = P_1(d_1) * P_2(d_2) * \cdots * P_n(d_n) = \sum_{d_1+d_2+\cdots d_n=D} \sum \cdots \sum \left[\prod_{i=1}^{n} P_i(d_i) \right] \tag{2}$$

where d_1, d_2, … , d_n are the durations of the subunits of word W, and $P_i(d_i)$ are the corresponding probabilities. To partially account for dependence between nearby phones we use context-dependent subword units, specifically triphones, for the purpose of estimating $P_W(d)$. The triphone duration distributions are estimated directly from the training corpus.

We used the ROS measure thus defined for all word tokens in the training data. We found that 80% of sentences with five or more words have both at least one word belonging to the fastest 33% and one word belonging to the slowest 33% of all words. This suggests that in conversational speech, speech rate is usually not uniform within a sentence.

Equation (1) can also be applied directly to subword units, thus allowing us to calculate the ROS of individual phones. This gives us an approach to study the variation of ROS at a more local level. For each word or sentence that has at least two phones, we computed the standard deviation of ROS of all of its phones. From the definition we see that the phone ROS ranges from 0 to 1; thus, its standard deviation must also fall within [0,1]. Dividing the interval [0,1] into 100 equivalent bins, we collected the histograms of phone ROS in both the within-word case and the within-sentence case on the whole training data, as depicted in Figure 2. The data suggests that the word is a better unit than the sentence for ROS modeling, since the average phone-level ROS deviation within a word is significantly smaller than within a sentence, which means that ROS is more stable at the word level than in the sentence level, and thus classifying each word as fast or slow makes more sense than classifying the entire sentence as fast or slow.

[Figure 2] ⇦

We note that the proposed ROS measure applies to individual words, and therefore does not include the duration of interword pauses that contribute to other common definitions of speech rate. The reason for this difference is that our approach aims at improving the acoustic modeling of the speech portions of the signal only, that is, the portions accounted for by word pronunciations. Furthermore, our goal is to do so by modeling the effects of ROS on the acoustic features, not by modeling ROS itself as a discriminator feature. For these reasons we are not concerned about ROS estimation *per se*, and have not investigated the quantitative relationship between our ROS measure and others proposed in the literature.

3. RATE-DEPENDENT ACOUSTIC MODELING

We focus on rate-dependent acoustic modeling alone, without changes to word pronunciations. In the proposed method, each word is given parallel pronunciations of “fast” and “slow” phone models. Both fast and slow pronunciations are initialized from the original rate-independent version, with a 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 and slow pronunciations are /w_f er_f d_f/ and /w_s er_s d_s/, consisting of fast and slow phone models, respectively. The recognizer automatically finds the pronunciations that maximize the likelihood score during 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 (but not within a word), thus accounting for speech rate variation within a sentence.

The introduction of parallel ROS-specific pronunciations is reminiscent of the introduction of parallel state paths in recent work on phone hidden Markov model (HMM) topologies (Iyer et al., 1999). Parallel path HMMs aim to model the acoustic consistency of adjacent frames, emulating trajectory-based segment models (Ostendorf et al., 1996), similar to the way our models enforce ROS consistency over adjacent phones.

3.1 Training rate-dependent acoustic models

Our initial experiments were performed on SRI's 1998 Hub-5 evaluation system (Weintraub et al., 1998), which uses continuous-density genonic HMMs (Digalakis et al., 1996) for acoustic modeling. The system used a multipass recognition strategy (Murveit et al., 1993). For the sake of simplicity, we ran our experiments with only the first-pass recognizer, which used gender-dependent non-crossword genonic HMMs (1,730 male genones and 1,458 female genones with 64 Gaussians per genone) and a bigram grammar with a 33,275-word vocabulary. The pronunciation dictionary was derived from the CMU version 0.4 lexicon with stress information stripped. Most words have a single entry in the CMU lexicon, while some have multiple entries for common pronunciation variants. For example, the word "was" has three entries: /w aa z/, /w ah z/, and /w ao z/. Generally speaking, the lexicon does not cover the possible pronunciation variations caused by different speaking rates. The recognizer used a two-pass (forward and backward) 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 acoustic features used were 9 Mel Frequency Cepstral Coefficients (C1-C8 plus C0) with their first- and second-order derivatives obtained from 18 filter banks covering 300-3300 Hz in 10 ms time frames. The acoustic training set consisted of 87 hours of speech for males and 106 hours for females, from a combination of corpora: (1) Macrophone read telephone speech, (2) 3,094 conversation sides from the BBN-segmented Switchboard-1 training set (with some hand-corrections), and (3) 100 CallHome English training conversations.

We first calculated the ROS for all words in the training corpus based on the above-mentioned measure, sorted words by ROS, and then split them into fast and slow categories. The ROS threshold for splitting was selected to achieve equal amounts of training data for fast and slow speech. The training transcriptions were labeled accordingly. We then prepared a special training lexicon: word tokens labeled "fast" were given pronunciations with fast phones only, and similarly for “slow” words. In this way, we were able to train fast and slow models simultaneously using the standard Baum-Welch training procedure.

We used the DECIPHER ^{TM} genonic training tools to run standard maximum likelihood estimation (MLE) gender-dependent training, and obtained rate-dependent models with 3,233 genones for male speech and 2,501 genones for female speech. The genone clustering for rate-dependent models used the same information loss threshold as previously used for rate-independent training.

Results. We compared the rate-dependent acoustic model with the rate-independent one (the baseline system) on a development subset of the 1998 Hub-5 evaluation data, consisting of 1,143 sentences from 20 speakers (9 male, 11 female). The first two rows of Table 1 show the WER for the two models. Note that all results reported here are based on speaker-independent within-word triphone acoustic models and a bigram language model, and are therefore not comparable to those for the full evaluation system.

[Table 1] ⇐

Rate-dependent modeling yields an absolute WER reduction of 1.9%, which was statistically significant ($p < 0.005$, using a matched pairs sign test). To eliminate the possible effect of different numbers of parameters, we adjusted the threshold in genone clustering to obtain a revised rate-independent model that had a number of parameters similar to that of the rate-dependent model. However, we did not observe any improvement from the increased number of parameters, suggesting that the improvement in the first experiment is indeed due to the modeling of ROS variation.

3.2 Bayesian adaptation versus standard training

In our previous work on the Broadcast News corpus (Zheng et al., 2000), instead of using the training method described above, we trained the rate-dependent model using a modified

Bayesian adaptation scheme (Digalakis and Neumeyer, 1996) 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 the earlier work, we had used a baseline system with very large models (256,000 Gaussians), and had classified the training data into three rate categories: fast, medium, and slow. For this model size, the training data was not sufficient to perform standard training. For the present Hub-5 task, by contrast, we had significantly more training data, and furthermore partitioned the data into only two classes instead of three, yielding more training data for each rate class. In addition, the optimal models with which we started were smaller. Thus, we were able to train the rate-dependent model robustly with standard training techniques, without resorting to adaptation. For comparison, we tested the Bayesian adaptation approach on the current training set. Similar to Zheng et al. (2000), while we used a separate rate-specific model 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.

Results. The last row of Table 1 shows the results on the same development data set as used earlier. We see that the adaptation-based ROS modeling approach brings an absolute win of 1.0% over the baseline, which is still statistically significant ($p<0.001$), but less of an improvement than the standard training scheme. This indicates that the difference between fast and slow speech in the acoustic space is significant, and that, given enough data, standard training might be better than the previously used adaptation scheme at capturing 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. The adaptation approach, on the other hand, does not recompute genonic clustering, does not change the transition probabilities, and used 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 perform as well as standard training.

3.3 Relation to explicit duration modeling

In the above method, although we used a ROS measure based on duration information to partition fast and slow speech for training ROS-dependent acoustic models, we did not explicitly model duration itself as a discriminator variable. Instead, we modeled the effects of ROS in the acoustic feature domain as a result of the different acoustic realizations under different speech rates. Explicit duration modeling, on the other hand, models duration directly as an observed variable, without necessarily considering the influence of duration on the acoustic features. While this latter approach is outside the scope of this paper, we have evidence to suggest that the two modeling approaches are complementary and additive in their benefits. As part of the 2000 and 2001 SRI Hub-5 recognition systems (Stolcke et al., 2000), we have combined the rate-dependent models described here with the explicit word and phone duration model of Gadde (2000). The explicit duration model gave an additional win over the rate-dependent acoustic model that was comparable to the win over a standard rate-independent model.

4. RATE-DEPENDENT PRONUNCIATION MODELING WITH ZERO-LENGTH PHONES

In the experiments described so far, we used identical pronunciations with different sets of phone models to account for the acoustic differences between fast and slow speech. However, it is well known that different speaking rates correlate with different pronunciations because of the different degrees of phone coarticulation and reduction (Fosler-Lussier and Morgan, 1998). In the work of Siegler and Stern (1995), a rule-based pronunciation dictionary transformation approach is used to produce specific pronunciations for fast speech. One major type of transformation used is the deletion of certain phones (especially, schwa) from dictionary entries under certain conditions specified by linguistic rules. We refer to this technique as the *phone-deletion approach*. This is a good modeling technique for dealing with phonetic elision, but may not be appropriate for modeling phonetic reduction, since it implicitly assumes that reduced phones have no acoustic evidence at all. Some level of modeling of reduction is still achieved at the acoustic model level; nevertheless, the minimal duration constraints of the standard HMM topologies may introduce errors in short realizations, typically for the reduced phones. Here, we propose a complementary approach to the modeling of phonetic reduction in fast speech, based on the notion of "zero-length phone" (Zheng et al., 2000). The idea is to allow certain short phone models to be skipped during the search, while preserving their contextual influences on the neighboring phone models. In this way some acoustic evidence of the reduced phones is preserved in the triphone models for the adjacent phones.

Let us examine the pronunciation of the word "bust" as an example to explain our approach and its advantages. In normal speaking mode, "bust" is pronounced /b ah s t/, whereas in fast speech, the /t/ might be too short to be described by a full phone model. Attempting to model this reduction with the phone deletion approach would result in an additional pronunciation for the word "bust", /b ah s/, which could be added to the dictionary to handle this case. However, this would introduce new confusability into the vocabulary, specifically, between the word "bust" and the word "bus". With zero-length phones, we instead add the pronunciation /b ah s t_Φ/ to the dictionary, where /t_Φ/ denotes a zero-length /t/. The advantage of this approach becomes clear once we examine which context-dependent models are assigned to the individual phones. We will use /a[b]c/ to denote an allophone of /b/ in the context /a_c/, where a or c can be phones or the symbol "#", denoting a word boundary. Thus, the reduced form /b ah s t_Φ/ would be realized by the triphone models /#[b]ah b[ah]s ah[s]t/. The realization of "bus", on the other hand, is /#[b]ah b[ah]s ah[s]#/, which is still different from that of the reduced form of "bust". Thus we introduce less interword confusability by using zero-length phones than by phone deletion, since we actually model the influence of the skipped phones. Another advantage of zero-length phones is that they introduce no new triphones. This point is important when pronunciations are generated automatically, resulting in numerous new pronunciations. The phone-deletion approach would add many new triphones in such cases, leading to more data fragmentation.

We developed an approach to generate reduced pronunciation variants for fast speech automatically, by letting a subset of the phones in a word's pronunciation be zero-length. We began by analyzing the duration distribution for each type of triphone based on forced Viterbi alignment of the training data with rate-independent acoustic models. We defined a pool of triphones that are candidates for realization as zero-length phones by using the following criteria: each triphone must have (1) at least 30 instances in the training corpus and (2) more than 35% probability of having a duration of three frames. Note that since we used three-state HMMs for all triphones, three frames is the minimum duration for a triphone. When the duration of a phone is less than three frames, the Viterbi alignment will have to "steal" frames from adjacent phones. In recognition, this problem could lead to poor acoustic match, and therefore, to recognition errors. If a large portion of the instances of a triphone have three-frame duration, it means that the real duration of the triphone could be less than three frames in many cases, which Viterbi alignment could not find. Then, based on forced Viterbi alignments, we collected pronunciation variants for each word having at least five instances in the training data. If a triphone in a pronunciation instance belonged to the pool of zero-length phone candidates, and its duration was exactly three frames, we converted this triphone to a zero-length phone, and added the resulting pronunciation for that word. To make this method robust with limited training data, we used both male and female data to obtain the pronunciation variants, and kept only those that occurred at least twice in the training corpus. Using this approach, we ended up selecting 8,465 new pronunciations that contain zero-length phones for 4,641 high-frequency words, covering 43.2% word tokens in the training corpus.

Using the pronunciation dictionary thus created, we trained rate-dependent acoustic models as described in Section 3, and assigned each pronunciation a probability based on its occupancy count. To avoid penalizing words with more than one pronunciation in the Viterbi search, pronunciation probabilities were scaled in such a way that the most frequently used pronunciation for a word has probability 1. Note that since we used distinct word tokens for fast and slow speech in training, the same pronunciation would have different probabilities in the fast and slow versions. Typically, the reduced pronunciation would have higher probability in the fast version than in the slow versions, and the opposite would be true for the full pronunciation. Thus, we effectively obtain rate-dependent pronunciation models.

To verify the effectiveness of the proposed zero-length phone treatment, we created another system based on the same set of pronunciations as above, but where highly reduced phones were deleted from the pronunciations, without introducing zero-length phones. We trained a separate set of acoustic models with the dictionary thus generated, and performed recognition on the same test data for comparison.

Results. The experiments with alternate dictionaries used as their baseline a recognition system developed for the March 2000 Hub-5 evaluation (Stolcke et al., 2000). The baseline system had been improved relative to the one in earlier experiments by using a wider-band front end (100-3760 Hz), higher-dimensional 39-component feature vectors, and vocal tract-length normalization (Wegmann et al., 1996). As shown in Table 2, the rate-independent baseline was thus improved by 5.2% absolute compared to the earlier system (cf. Table 1). Relative to the improved baseline, the rate-dependent system without changes to the dictionary now yields a WER reduction of 1.9% absolute (3.5% relative), on a par with the improvement in the earlier system. An additional 0.3% absolute improvement was achieved by also including rate-dependent pronunciations based on zero-length phones. The overall improvement over the rate-independent baseline was therefore 2.2% absolute (or 5.4% relative). The improvement of the two rate-dependent systems compared to the baseline is statistically significant ($p<0.001$), although the difference between them does not reach significance ($p<0.2$).

The result from the control experiment (last row of Table 2) confirms the importance of our zero-length-phone treatment: With the phone deletion approach, we did not obtain any win from adding “fast” pronunciations; in fact, the WER increased slightly (0.2% absolute). We take this as evidence that the zero-length phone approach gives a more accurate model of phonetic reduction with fewer side effects of increased confusability and fragmentation of the triphone space than the phone-deletion approach.

[Table 2]

5. ROS MODELING FOR MULTIWORDS

In previous experiments, we applied ROS modeling at the word level and obtained improvements over the rate-independent baseline. As described here, we also investigated the interaction of rate-dependent acoustic modeling with the popular “multiword” pronunciation modeling technique. This technique has proven highly successful in conversational speech recognition, thus giving us an even more competitive baseline against which to test our ROS modeling approach. These experiments will provide further evidence for our choice of words as the appropriate units for modeling ROS variation.

A multiword is a high-frequency word N-gram, such as “going to”, that is handled as a single unit in the vocabulary, typically with some specific pronunciations in the dictionary. Like other researchers we found that introducing multiword modeling can lead to substantial WER reduction in conversational speech recognition (Finke and Waibel, 1997; Stolcke et al., 2000).

Multiwords improve recognition in three ways. First, based on linguistic knowledge or machine learning techniques, new phonetic pronunciations are introduced that account for various elision, reduction, and coarticulation phenomena at both phone and syllable levels. We will illustrate this by taking the bigram “going to” as an example. In the initial dictionary, “going” has one pronunciation: /g ow ax ng/; “to” has two pronunciations: /t uw/ and /t ax/. For “going to”, in addition to simply integrating the two words’ pronunciations, in this case /g ow ax ng t uw/ and /g ow ax ng t ax/, we manually added two extra entries: /g ow ax n t ax/ and /g ah n ax/. In fact, we found that /g ah n ax/ occurs with higher frequency than the other pronunciations, especially in fast speech. These specially introduced pronunciations are highly context dependent, and thus add little interword confusability. For example, in the pronunciation /g ah n ax/, “going” reduces to /g ah n/, and “to” to /ax/; however, this variant is restricted to the compound "going to", and will not affect other cases. Second, a multiword is treated as a single word unit in the dictionary; thus allowing within-word triphones to be used in modeling cross-word coarticulation effects, as if cross-word triphones had been used (but without the computational overhead of general cross-word context-dependent phone modeling). Third, the existence of multiwords effectively extends the range of the language model: a bigram of multiwords can span up to a 6-gram of regular words.

Results. We applied the rate-dependent modeling technique of Section 3, that is, using rate-specific phone models for rate-specific pronunciations, to a multiword-augmented version of the recognition system described in Section 4. We added 1,389 word bigrams and trigrams to the dictionary as multiwords, covering about 42% of the word tokens in the training data. The multiword-based system yielded a baseline WER that was reduced almost 10% relative to the system without multiwords (cf. the first rows in Tables 2 and 3).

For several reasons we did not include reduced pronunciations with zero-length phones in this experiment. First, multiwords already capture many of the phonetic reductions occurring in frequent (especially function) words. Second, the multiword-augmented dictionary was already quite large, and further adding to it with reduced pronunciations would have negatively impacted recognition speed, without a large reduction in WER to be expected based on the earlier results. Note that even without zero-length phones, we used rate-dependent pronunciation probabilities for both multiwords and regular words.

In the first experiment we treated multiwords as ordinary words, so no rate switching was allowed inside multiwords. As shown in the first two rows of Table 3, rate-dependent modeling yielded a WER reduction of 0.5% absolute, considerably smaller than in the system without multiwords (cf. Table 2), and not statistically significant ($p<0.2$).

[Table 3] ⇦

The diminished effectiveness of ROS modeling can be attributed to two factors. One reason is that each multiword had both standard (from the CMU dictionary) and reduced pronunciations, thereby accounting for ROS variation to some extent already. A second reason is that we failed to model the possible rate changes within multiwords. Figure 3 histograms the standard deviation of phone-level ROS over single words, multiwords, and whole sentences. The figure shows that ROS varies considerably more within multiwords than within regular words, highlighting the need to allow ROS changes within multiwords.

To address the problem of ROS change within multiwords, we tested two alternative schemes. In the first approach, separate rate-independent phonetic models were used to describe multiwords exclusively. Accordingly, we trained three classes of phone models: fast models for ordinary words, slow models for ordinary words, and a separate set of phone models trained only on the multiword data.

In the second scheme, we allowed rate switching at word boundaries within multiwords. To do so, we had to reintroduce word boundaries within our multiword pronunciations, which is not always obvious in the case of greatly reduced forms. We did so in a way that roughly minimizes the overall number of distinct word pronunciations. For example, we placed a word boundary (and thus a potential change point for ROS) before /ax/ in the form /g ah n ax/ of the multiword "going_to", since the word "to" has /ax/ as a reduced form in multiple contexts.

As shown in the last two rows of Table 3, we see improvements with both schemes (the difference relative to the baseline is statistically significant in both cases, $p<0.05$ and $p<0.001$, respectively). As might be expected, allowing rate switching within multiwords led to the best result, reducing WER by 1.5% absolute (3.0% relative) compared to the baseline rate-independent multiword system. To keep the comparison fair, we deliberately used the same number of Gaussians in the best-performing rate-dependent system as in the rate-independent baseline system. The relative WER reduction is similar to that in the system without multiwords, showing that rate-dependent acoustic modeling is complementary to, and nearly additive with, the improved pronunciation modeling afforded by multiwords. Furthermore, since rate switching is apparently frequent within multiwords, and especially since multiwords consist mostly of short function words, our results confirm the choice of word units as the appropriate locus of modeling for local ROS change, in agreement with Figure 3.

[Figure 3] ⇦

6. CONCLUSIONS

We have proposed a rate-dependent acoustic modeling scheme that accounts for within-sentence speech rate variation and does not rely on ROS estimation prior to recognition. Experiments show that this method results in 3.0-3.5% relative word error rate reduction on conversational telephone speech, using a variety of baseline systems. Acoustic ROS modeling combines well with, and is additive with, independent improvements in pronunciation modeling based on an extensive use of multiwords. It proved important to allow rate changes at word boundaries within multiwords, underlining the importance of local ROS variation in conversational speech.

We also developed a novel approach for modeling phone reductions common in spontaneous and fast speech. We showed that it was better to model highly reduced phones with “zero-length” phones, thus preserving the phone contexts in adjacent phones, rather than simply deleting phones from the pronunciations. The latter approach was found to degrade recognition accuracy, while the zero-length phone approach yielded a small (0.3% absolute) WER reduction in addition to rate-dependent acoustic modeling.

The work so far can be extended in several directions. First, our rate-dependent acoustic modeling approach can be generalized to use a finer-grained measure of ROS instead of a simple “fast” versus “slow” distinction. Eventually, it would be desirable to entirely eliminate discrete ROS classes, for example, by directly using a continuous ROS feature in decision tree clustering of acoustic models (Paul, 1997). The probabilistic density functions for different context-dependent phones would thus be tied according to the actual influence of ROS on their acoustic realization, mitigating the data fragmentation brought about by fixed ROS classes.[4] In addition, a special language model can be used to model rate switching probabilities.

Second, other parts of the acoustic modeling architecture (besides acoustic feature distributions and pronunciations) could be conditioned on speaking rate. In particular, the topologies of HMMs could vary depending on the level of ROS. For example, "fast" phone models could use fewer states (and thus shorter minimum durations) than "slow" phone models. ROS-specific HMM topologies are straightforward to accommodate in our modeling approach since conditioning on ROS is already implemented by separate phone models.

Third, the current zero-length phone approach models mainly highly reduced realizations. We plan to combine it with the phone-deletion approach, which is preferable for modeling phone elision phenomena, to provide a continuum of methods to model different levels of pronunciation variation. In addition, we will introduce more phonetic knowledge in the automatic procedure of pronunciation variant generation, to improve the quality of the resulting dictionary.

ACKNOWLEDGMENTS

We thank Colleen Richey for assistance with dictionary creation, and Ramana Rao for his contributions to the recognition system. We thank Dr. Guy Perennou for his detailed and instructive comments in the reviews, and for providing the French translation of the abstract.

REFERENCES

Digalakis, V., Monaco, P., and Murveit, H. (1996), "Genones, generalized mixture tying in continuous hidden Markov model-based speech recognizers," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 4, no. 4, pp. 281-289.

[4] Decision-tree clustering according to ROS was suggested by one of the anonymous reviewers. Our present acoustic model training relies on bottom-up clustering (Digalakis et al., 1996) and therefore forced us to use discrete ROS classes.

Digalakis, V., and Neumeyer, L. G. (1996), “Speaker adaptation using combined transformation and Bayesian methods,” *IEEE Trans. Acoust. Speech Signal Process.*, vol. 4, no. 4, pp. 294-300.

Finke, M., and Waibel, A. (1997), “Speaking mode dependent pronunciation modeling in large vocabulary conversational speech recognition,” *Proc. EUROSPEECH*, vol. 5, pp. 2379-2382, Rhodes, Greece.

Fosler-Lussier, E., and Morgan, N. (1998), “Effects of speaking rate and word frequency on pronunciations in conversational speech,” *Speech Communication*, vol. 29, no. 2-4, pp. 137-158.

Gadde, V. R. R. (2000), “Modeling word duration for better speech recognition,” *Proc. NIST Speech Transcription Workshop*, College Park, MD.

Gonzalez, R., and Woods, R. (1992), *Digital Image Processing*, Addison-Wesley Publishing Company, Chap. 4.

Iyer, R., Kimball, O., and Gish, H. (1999), “Modeling trajectories in the HMM framework,” *Proc. EUROSPEECH*, vol. 1, pp. 479-482, Budapest.

Mirghafori, N., Fosler E., and Morgan N. (1996), “Towards robustness to fast speech in ASR,” *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process.*, vol. 1, pp. 335-338, Atlanta.

Morgan, N., and Fosler, E. (1998), “Combining multiple estimators of speaking rate,” *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process.*, vol. 2, pp. 729-732, Seattle.

Murveit, H., Butzberger, J., Digalakis, V., and Weintraub, M. (1993), “Large-vocabulary dictation using SRI’s DECIPHER(TM) speech recognition system: Progressive-search techniques,” *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process.*, vol. 2, pp. 319-322, Minneapolis.

Ostendorf, M., Digalakis, V. V., and Kimball, O. A. (1996), “From HMM’s to segment models: A unified view of stochastic segment modeling for speech recognition,” *IEEE*

Trans. Speech and Audio Process., vol. 4, no. 5, pp. 360-378.

Paul, D. (1997), "Extensions to phone-state decision-tree clustering: Single tree and tagged clustering," *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process.,* vol. 2, pp. 1487-1490, Munich.

Richardson, M., Hwang, M., Acero, A., and Huang, X. D. (1999), "Improvements on speech recognition for fast talkers," *Proc. European Conf. Speech Communication Technology*, vol. 1, pp. 411-414.

Siegler, M.A., and Stern, R.M. (1995), "On the effects of speech rate in large vocabulary speech recognition systems," *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process.*, vol. 1, pp. 612-615, Detroit.

Tuerk, A., and Young, S. (1999), "Modelling speaking rate using a between frame distance metric," *Proc. European Conf. Speech Communication Technology*, vol. 1, pp. 419-422, Budapest.

Stolcke, A., Bratt, H., Butzberger, J., Franco, H., Gadde, V. R. R., Plauché, M., Rickey, C., Shriberg, E., Sönmez, K., Weng, F., and Zheng, J. (2000), "The SRI March 2000 Hub-5 conversational speech transcription system," *Proc. NIST Speech Transcription Workshop*, College Park, MD.

Wegmann, S., McAllaster, D., Orloff, J., and Peskin, B. (1996), "Speaker normalization on conversational telephone speech," *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process.,* vol. 1, pp. 339-341, Atlanta.

Weintraub, M., et al. (1998), "SRI System Description," Ninth Hub-5 Conversational Speech Recognition Workshop, Linthicum Heights, MD.

Zheng, J., Franco, H., Weng, F., Sankar, A., and Bratt, H. (2000), "Word-level rate-of-speech modeling using rate-specific phones and pronunciations," *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process.,* vol. 3, pp. 1775-1778, Istanbul.

	Male	Female	All
Rate-independent model	55.3	63.4	59.8
Rate-dependent model from training	52.9	61.9	57.9
Rate-dependent model from adaptation	54.0	62.6	58.8

Table 1: Comparison among the baseline (rate-independent) model, the rate-dependent model from standard training, and the rate-dependent model from adaptation (in % WER on the development set).

	Male	Female	All
Rate-independent system	50.6	57.9	54.6
System with rate-dependent phones	49.2	55.6	52.7
System with both rate-dependent phones and pronunciations (zero-length-phone treatment)	48.5	55.6	52.4
System with both rate-dependent phones and pronunciations (phone-deletion treatment)	48.7	56.2	52.9

Table 2: WER of a rate-independent baseline system, a system with rate-dependent phones, a system with both rate-dependent phones and pronunciations, where pronunciations were treated with zero-length phones, and a similar system with pronunciations treated by the common phone deletion scheme.

	Male	Female	All
RI baseline multiword system	44.3	53.3	49.3
RD system, no rate change within multiwords	43.6	53.0	48.8
RD system with multiword-specific RI models	43.6	52.6	48.6
RD system, allow rate change within multiwords	42.6	51.9	47.8

Table 3: WER results using different schemes of rate-dependent (RD) modeling compared to a rate-independent (RI) baseline system containing multiwords.

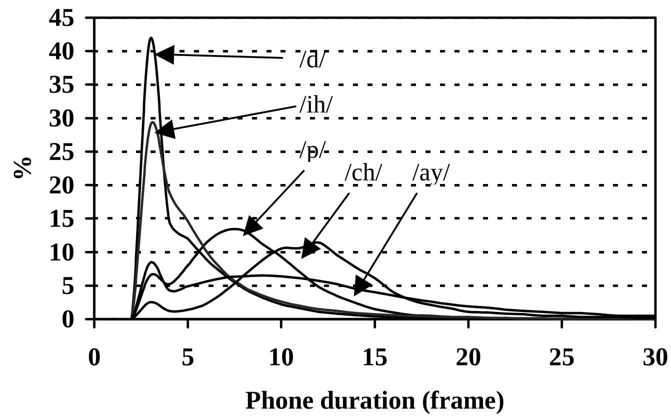

Figure 1: Duration distributions of phonemes /d/, /p/, /ch/, /ih/ and /ay/.

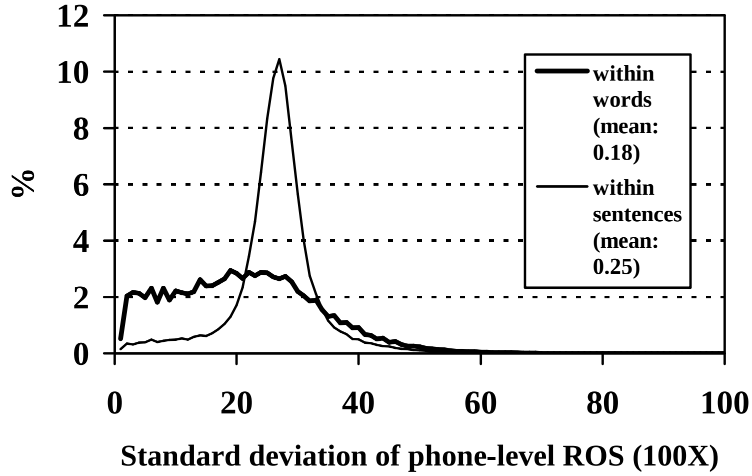

Figure 2: Histograms of the standard deviation of phone-level ROS within words vs. within sentences. The X-axis is the standard deviation of phone-level ROS multiplied by 100.

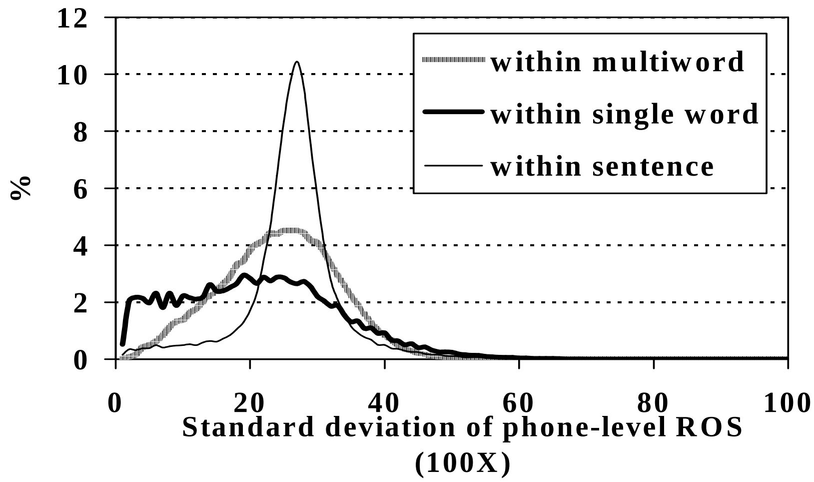

Figure 3: Histograms of the standard deviation of phone-level ROS within multiwords, within single words, and within sentences. The X-axis is the standard deviation of phone-level ROS multiplied by 100.