MULTIRATE ASR MODELS FOR PHONE-CLASS DEPENDENT N-BEST LIST RESCORING

Venkata R. Gadde *Kemal Sönmez* *Horacio Franco*

Speech Technology and Research Laboratory
SRI International, Menlo Park, CA, USA
{rao,kemal,hef}@speech.sri.com

ABSTRACT

Speech comprises a variety of acoustical phenomena occurring at differing rates. Fixed-rate ASR systems assume in effect a constant temporal rate of information flow via incorporating uniform statistics in proportion to a sound's duration. The usual tradeoff window length of 25-30 milliseconds represents a time-frequency resolution compromise, which aims to allow reasonable speed for following changes in the spectral trajectories and sufficient number of samples to estimate the harmonic structure. In this work, we describe a technique to augment a recognizer that uses this compromise with information from multiple-rate spectral models that emphasize either better time or better frequency resolution in order to improve performance. The main idea is to use the hypotheses generated by a fixed-rate recognizer to determine the appropriate model rate for a segment of the speech waveform. This is realized through a technique based on rescoring of N-best lists with acoustical models using different temporal windows by a phone-dependent posterior-like score. We report results on the NIST Evaluation 2002 dataset, and demonstrate that the rescoring method produces word error rate (WER) improvements in a baseline system.

1. INTRODUCTION

Modeling of speech with fixed rate window front-end hidden Markov models implies a constant rate of information accumulation. In this framework, frames of a fixed length are scored uniformly to compute the likelihood that a given sequence of feature vectors is produced by the model. The common fixed frame length of 25-30ms represents a fundamental time-frequency trade-off in the speech representation. For example, vowels can result in a relatively stationary harmonic

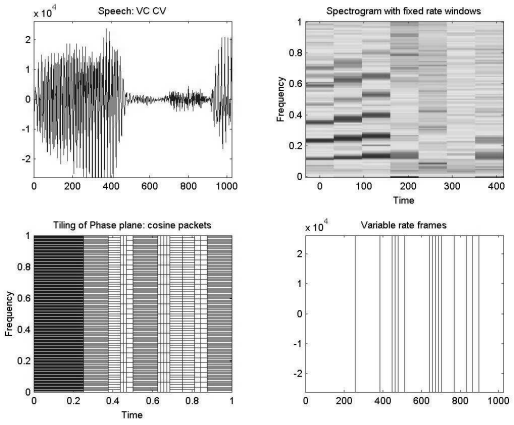

Figure 1: Variable frame rate by cosine packet decomposition. Top left: waveform. Top right: spectrogram with a fixed window length. Bottom left: signal adaptive (best basis with cosine packets) time-frequency tiling. Bottom right: variable rate frames.

structure that can be sustained for hundreds of milliseconds, whereas stop consonants can have landmark transients that last no more than ten milliseconds. It is well known that a frame length of 25-30 ms is too long for capturing information-bearing transient phenomena which may have durations as short as a couple of ms. At the same time, stationary segments have constant spectral characteristics for much longer, on the order of 100ms. These observations motivate exploring techniques that zoom in temporally on transient events, and zoom out temporally on stationary events for acoustical modeling with a finer frequency resolution. In a constant frame length frontend, transient phenomena are blended with the context, decreasing the sharpness of the models that account for information-bearing discontinuities. Therefore, frame scores with particularly relevant

information, such as those of stops, are washed out in the statistics of phone scores that span many more frames, such as those of vowels. In [8], we made an assessment of the extra information by exploring a set of acoustic phenomena that discriminate a class of easily-confused consonants, stops in vocalic contexts. On a collected read-speech database [6] and on Switchboard, we demonstrated that a small set of features computed with the right time-frequency resolution discriminate voiceless stops in CV contexts successfully.

Incorporation of information from acoustical phenomena taking place at different rates has received significant amount of attention in the ASR literature. In particular, there are two main classes of approaches that are related closely to the approach we are proposing in this work. The first class of models [2,3,4,5,7] attempts to address the issue of varying time-frequency tradeoff of speech at the very basic representation level in the frontend of an ASR engine. The speech signal is represented by a sequence of frames at variable rates. In previous work [8], we proposed cosine packets to generate signal-adaptive tilings of the time-frequency plane, which can be used to generate variable rate of frames (see Figure 1) In [9, 10], an alternative method for generating variable rate frames through the definition of a metric that measures the rate of change in the waveform using the entropy of the Gaussian model distribution. By allowing frames to extend to variable temporal windows, the score contributions from phones with varying durations are balanced, and the appropriate time-frequency tradeoff is selected on the fly.

The second class of related approaches is more general and aims to incorporate diverse information sources at the model level by generalizing HMMs through allowing multiple streams. In [1], graphical models are used to combine information from multiple models addressing different time-frequency tradeoffs. Different sets of features aiming to model fast and slow varying parts of the acoustic representation generate streams that form the structure of a dynamic Bayesian network. In this type of approach, the selection of the time-frequency trade-off is pushed back to the model level, and graphical model parameters determine how information sources stemming from different resolutions are combined.

In this work, we are presenting a method that aims to incorporate information from multiple time-frequency tradeoffs through projecting the variable frame problem at the frontend to the backend, i.e. through rescoring of N-best lists generated by a fixed-rate recognizer with a normalized rate-dependent score. In our approach, the hypotheses generated by the fixed-rate ASR engine are in effect used to parse the incoming speech into phones, which subsequently determine the most likely rate model through the definition of a mapping from phone classes to the available set of multiple rate models. The method involves rescoring of N-best lists, a common way of incorporating other information sources, by phone-dependent multiple rate model scores. The scoring has two important aspects that differentiate our approach. One important novel aspect of the scoring is the normalization with respect to dynamic range of scores of models at different rates, which is carried out by normalizing with the likelihood of all the phones in the same phone context at the same resolution. Another important aspect is the averaging of the frame-level scores to produce a single score for each phone-state in the hypotheses. Resulting phone-class dependent scores are treated as knowledge sources and combined into a linear model, parameters of which is optimized to minimize the WER.

The paper is organized as follows. In Section 2, we describe the proposed phone-dependent rescoring method along with normalization of the phone-dependent scores. In Section 3, we detail the basic infrastructure and the experiments carried out, and present and discuss our results in Sections 4 and 5, respectively.

2. **APPROACH**

The main signal-adaptive rate modeling approach we present in this paper involves N-best rescoring using acoustic models trained at different rates. Rate estimation in this case is not directly signal dependent but is tied to the phone class, and can efficiently be integrated into a fixed-rate recognizer. Since the likelihood scores from multirate models are not directly comparable, we use log-likelihood ratios for frame-level scores, and compute the average of all frame-level scores for a given state.

In this approach, adaptation of the rate with respect to the signal is indirect through the hypotheses of a fixed-rate recognizer. There are advantages and disadvantages associated with this type of recognizer back-end framework. The performance of the approach is bounded with the 2000-best hypotheses in the original hypotheses list, but it also benefits from phone-dependent information to which a signal-adaptive tiling such as the best basis algorithm cannot have access. This approach also requires significantly less implementation to experiment with and has far fewer parameters to tune than the directly signal dependent rate modeling shown in Figure 1.

2.1. Phone-class dependent N-best list rescoring with multirate models

The technique involves choosing a small set of rates and training acoustic models at those rates. After generating N-best lists using standard rate model, we score the N-best hypotheses using different rate models and combine the scores to minimize the word error rate.

The steps of the technique can be summarized as follows:

1. Choose a set of rates (0.67, 1.0, and 2.0 times the fixed-rate) and train acoustic models at those rates.
2. Generate a map from phone classes to the set of rates (this step is omitted in most of our experiments – see explanation below).
3. Generate N-best lists using a baseline acoustic model (i.e., 100 frames/sec rate model)
4. Generate phone-class dependent scores for each phone in the hypotheses
5. Rescore the N-best hypotheses to generate phone-class dependent hypothesis scores as additional knowledge sources
6. Combine all the scores to minimize the word error rate over the N-bets list.

In the following experiments, instead of pre-assigning the rates to different phone classes, we scored each phone-state using all rate models. These scores are input to the combiner and that the combiner determined appropriate weights.

2.2. Frontend parameters

For the multiple rate models there are two fundamental front-end parameters that need to be adjusted: the frame shift, and the window size. In this work, we tie both parameters proportionally to the selected rates.

Model	Frame shift	Window size
Baseline (1.0)	10 ms	25.6 ms
Baseline x0.67	15 ms	38.4 ms
Baseline x 2	5 ms	12.8 ms

Table 1: Frontend parameters for the multirate models.

2.3. Normalization for phone-class dependent scores

The main issue in combining scores across heterogeneous models is that the likelihood scores are not comparable across different rate models, and therefore can not directly be combined into a parametric functional form to minimize the WER.

The solution we propose in this work is to use a normalized phone-class likelihood ratio for frame-level scores. Specifically, the normalized score for feature vector x_i at triphone-state ($p_p p p_n$) is computed by

$$\hat{S}(x_i, p) = S(x_i \mid p_p p p_n) = \log\left[\frac{P(x_i \mid p_p p p_n)}{\sum_{j \in Phone-classes} P(x_i \mid p_p p_j p_n)}\right] \quad (1)$$

where p represents the center phone and p_p and p_n represent the preceding and succeeding phone contexts.

Given the normalization in Eq. (1), each frame score $\hat{S}(x_i, p)$ can now be regarded as independent of the rate of the model by which it was generated. With the normalized scores, it is now possible to accumulate sentence level scores for phone classes to be combined with weights as explained in the next section.

2.4 Phone-class dependent scoring and score combination

Rates are assigned to different phone classes through exploration of phone-class clustering experiments starting

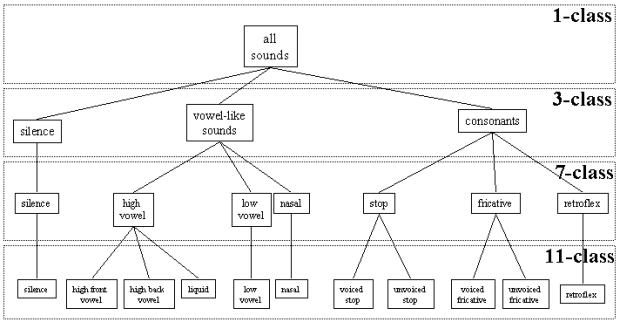

Figure 2: Phone class sets used in experiments.

from an a priori mapping based on phone characteristics. Fig. 2 shows the clustering among the different phone classes that we used in our experiments. The selected mapping can be used during rescoring to select the appropriate rate model to score the states belonging to a given phone class. This results in the generation of sentence-level scores for different phone classes, *Pk*, as given by Eq. (2).

$$\tilde{S}(P_k) = \sum_i \hat{S}(x_i, p) I[p \in P_k] \quad (2)$$

Finally, we combine the phone-class dependent scores with the baseline acoustic and language model scores though a linear combiner and optimize the linear combiner weights to directly minimize the WER as shown in Eq. (3).

$$S_{combined} = \beta_0 S_{AM} + \beta_1 S_{LM} + \sum_k \alpha_k \tilde{S}(P_k) \quad (3)$$

3. EXPERIMENTS

In our experiments, multiple rate modeling involved the training of the following three models:

1. Baseline model at standard rate (100fps, 25.6ms window)
2. Model at rate 0.667 (15ms shift, 38.4ms window)
3. Model at rate 2.0 (5ms shift, 12.8ms window).

The acoustic training data were the male subset of our RT02 CTS training set (~140 hours). The features were 13 mel frequency cepstra (including C0) and their first and second time derivatives. Feature normalizations included speaker level vocal tract length normalization, and z-norm. We did not use HLDA or CMLLR (SAT) in this set of experiments. We trained non-cross word triphone models containing 2400 genones (state clusters) with 64 gaussians per genone. The models were adapted to a phone-loop (using MLLR) before recognition. Two test sets were used for experiments: (i) A subset of the dev2001 male set (863 utterances containing 10328 words), used for both tuning and testing, and (ii) the eval2002 (RT02) male set (3083 utterances containing 36073 words). The baseline model was used to generate 2000-best lists of hypotheses using a bigram language model. The 2000-best lists were rescored with each rate model. Finally, the scores were combined and the optimal weights for combination were estimated to arrive at the WER computed using the best hypotheses.

4. RESULTS

To examine the relative performances of the three rate models, we did a recognition experiment on the dev2001 set. We used a bigram language model in this experiment. Table 1 shows the recognition results. We observed that the standard front-end has the lowest word error rate (WER). The lower frame rate model was a little worse but the faster frame rate model was significantly worse.

Table 2 shows the results of combining different rate models using N-best rescoring approach. We observed on the Dev01 test set that all rate models give reasonable improvements with the best reduction in WER 0.9% absolute for the rate_2.0 model. The combination of two rate models with the baseline scores gives a bigger improvement, 1.4% absolute. As the dev2001 set was used for both tuning and testing, we did another set of combination experiments on the Eval02 set. This set was partitioned into two parts, a tuning set containing about 1400 utterances and a held out set containing the rest. The tuning set was used to optimize the weights for different scores. These weights were then applied to the held out set and the WER was estimated. Table 3 shows the results from different model combinations. We show the results separately for the tuning and held out sets.

The results on Eval02 confirm that our rescoring approach results in a significant reduction in the WER, with the best reduction of 1.0% absolute for the rate_0.67 model. Increasing the number of classes reduces the WER for the tuning set but not on the held out set. For finer phone sets, it seems that we may need a larger tuning set to properly estimate the weights.

In performing experiments involving optimization of a large number of weights (18 weights for the 3-model comb. For 7-classes), we observed that the initial choice of weights affects the final outcome. To overcome this, we found that we need to first estimate the weights for the smaller number of classes first and use those weights as initial weights for the larger number of classes. It is possible that we could reduce the number of weights to be estimated by choosing a single appropriate rate for each phone class instead of using all models for all rates. In one experiment, we partitioned the phone classes into two groups, the first group containing slowly changing sounds (vowels and vowel-like sounds) and another group containing more rapidly changing sounds. Table 4 shows the phone class to rate map for the 7-class case. It can be seen that the slowly varying sounds were scored using the slower rate models and the faster varying sounds were scored using the fast rate models. The baseline model was used to rescore all sounds.

Table 5 shows the comparison of the results of this experiment with the experiment where all phones were rescored using all the different rate models. It is interesting to note that the WERs obtained are the same but the weights estimated were reduced by a third.

Table 1: Results on the NIST Dev2001 dataset

Model	WER
Baseline	37.0
Rate 2.0	49.8
Rate 0.67	39.0

Table 2: Results with phone-class-based rescoring on the NIST Dev2001 dataset

Model	WER for #Phone Class			
	1	3	7	11
Baseline	36.7	-	-	-
+0.67	36.3	36.4	36.3	36.2
+2.0	36.0	36.1	35.8	35.9
+0.67 +2.0	36.0	35.9	35.3	35.4

Table 3: Results with phone-class-based rescoring on the NIST Eval2002 dataset

Model	WER for #Phone Classes					
	1		3		7	
	Tune	Held	Tune	Held	Tune	Held
Baseline	39.9	39.9				
+2.0	39.2	39.4	38.8	39.1	38.6	39.2
+0.67	39.3	39.1	38.9	39.0	38.9	38.9
+0.67 +2.0	39.3	39.4	38.7	39.1	38.6	38.9

Table 4: Selection of phone classes for scoring with different rate models.

Rate	Classes Scored
2.0	Stop, fricative, retroflex, silence
0.67	High vowel, low vowel, nasal

Table 5: Results of selective scoring of different phone classes at different rates

Rescoring method	No. of weights	WER
Rescore all classes	18	38.9
Rescore selected classes	11	38.9

5. DISCUSSION AND FUTURE WORK

In conclusion, we proposed a multirate modeling approach based on rescoring of N-best lists with multiple rate models as an alternative to signal-adaptive variable rate modeling. The technique relies on phone-dependent rate models that are used to rescore the N-best lists of a fixed-rate recognizer. We introduced a novel normalization technique that allows normalized scores from multiple rate models to be integrated into a combined score directly to reduce WER. We obtained encouraging results from the N-best rescoring approach. Combining with various rate scores seems to help the system performance. In contrast, rescoring by any other acoustic model, such as rescoring the MFCC N-best with the PLP CW models, did not produce any improvement, therefore the gains are not solely due to arbitrary combination of systems. There remain issues to be investigated regarding the optimization of system weights, for example, for rescoring with systems that use crossword models.

6. REFERENCES

1. O. Cetin and M. Ostendorf, “Cross-stream Observation Dependencies for Multi-stream Speech Recognition,” Proc. Eurospeech, pp. 2517–2520, 2003.
2. J. J. Hant and A. Alwan, “A psychoacoustic masking model to predict the perception of speech-like stimuli in noise,” Speech Communication, vol. 40, pp. 291–313, May 2003.
3. P. Le Cerf and D. Van Compernolle, “A new variable frame rate analysis method for speech recognition,” IEEE Signal Processing Letter, vol. 1, no. 12, pp.185–187, December 1994.
4. J. Luettin, G. Potamianos and C. Neti, “Asynchronous stream modeling for large vocabulary audio-visual speech recognition,” Proc. ICASSP, pp. 169-172, 2001.
5. J. Macias-Guarasa, J. Ordonez, et al., “Revisiting scenarios and methods for variable frame rate analysis in automatic speech recognition,” Proc. Eurospeech, pp. 1809–1812, 2003.
6. M. Plauche, M. K. Sönmez, ``Machine Learning Techniques for Identification of Cues for Stop Place", *Proc. of ICSLP 2000*, Beijing, China, October 2000.
7. K.M. Pointing and S.M. Peeling, “The use of variable frame rate analysis in speech recognition,” Computer Speech and Language, vol. 5, no. 2, pp. 169–179, April 1991.
8. K. Sonmez, M. Plauche, and E. Shriberg, "Consonant discrimination in elicited and spontaneous speech: A case for signal-adaptive front ends in ASR," in Proc. ICSLP, vol. 1, (Beijing), pp. 325--329, October 2000.
9. H. You, Q. Zhu, and A. Alwan, “Entropy-based Frame Rate Analysis of Speech Signals and its Application to ASR,” Proc. IEEE ICASSP, pp. 549–552, 2004
10. Q. Zhu and A. Alwan, “On the use of variable frame rate analysis in speech recognition,” ICASSP, pp. 3264–3267, 2000.