AN INVESTIGATION OF VARIABLE BLOCK LENGTH METHODS FOR CALCULATION OF SPECTRAL/TEMPORAL FEATURES FOR AUTOMATIC SPEECH RECOGNITION

Montri Karnjanadecha[*] *and Stephen A. Zahorian*[**]

[*]Department of Computer Engineering, Faculty of Engineering
Prince of Songkla University, Hatyai, Songkla, Thailand 90112

[**]Department of Electrical and Computer Engineering
Old Dominion University, Norfolk, VA, 23529 USA

ABSTRACT

This paper presents an investigation of non-uniform time sampling methods for spectral/temporal feature extraction for use in automatic speech recognition. In most current methods for signal modeling of speech information, "dynamic" features are determined from frame-based parameters using a fixed time sampling, i.e., fixed block length and fixed block spacing. This work explores new methods in which block length and/or block spacing are variable. Three methods are suggested and each was tested with the TIMIT database using a standard HMM recognizer. Phone recognition experiments were conducted using the standard 39 phone set. The methods were also evaluated with various HMM model complexities. Experimental results indicated that none of the proposed non-uniform feature time sampling methods perform significantly better than fixed time sampling methods. However, the best results obtained with the front end are comparable to those obtained with current state-of-the-art systems. Also the performance of our monophone system surpasses that of most reported context-dependent monophone systems.

1. INTRODUCTION

A standard strategy to compute spectral/temporal features for automatic speech recognition uses uniform temporal resolution with respect to spectral features. That is, features are computed using a fixed frame length and spacing for frame-based parameterization and fixed block (segment) length and spacing for feature trajectory calculations. The underlying assumption is that speech information is uniformly distributed in time across an utterance. However, both the physics underlying speech production and available psychophysical evidence on speech perception contradict this time uniformity assumption. For example, vowels are the longest phones in English with relatively fixed spectra whereas stop consonants are much shorter and more dynamic in nature. The informational importance of stops is generally considered to be at least as great as that of vowels and humans are also quite proficient at recognizing stops. Although a statistical modeling framework such as Hidden Markov Model (HMM) can account for duration variability, still HMMs implicitly "score" observations proportional to length. Thus, an HMM recognizer typically performs similarly to human performance for the long duration vowels , but is more degraded relative to human performance for the short duration stops. An HMM recognizer also has difficulty with tasks such as the "e" set, where most of the distinguishing characteristics occur over a small time portion of each token. Another consideration is that the large degree of redundancy in the long steady sounds could degrade the performance of an HMM since this situation contradicts the HMM assumption that successive observation vectors are independent.

Motivated by some of the comments above, this paper presents some techniques that compute spectral/temporal parameters with non-uniform time sampling. After frame features have been calculated, a Discrete Cosine Transform (DCT) over time was used to encode blocks of frame features. Note this approach to calculating feature trajectories is not the same as the linear regression approach more typically used. The primary goal of this paper is to investigate methods which emphasize the temporal characteristics of speech in the areas where the spectrum changes rapidly, while de-emphasizing those areas where spectrum changes slowly.

Temporal information can be better captured when measured with a short time widow. This corresponds to the use of a short block length to encode the information or to closely sample the data in that area. Based on this basic notion, a method called 'variable block length' was explored and tested. This method extracts the spectral/temporal features using blocks of variable length. The length of a block was determined by minimizing a local reconstruction error. The second method used a fixed block length fixed but non-uniformly advanced the block depending on the magnitude of a spectral derivative measure. This method is called 'variable block spacing'. The last method tested utilized both short block and long block to capture rapid spectral changes and slow spectral changes. Principal Components Analysis (PCA) was used to reduce the dimensionality of the combined features.

The methods were tested with the TIMIT database using the NIST training set and the NIST core test set. An HMM recognizer was constructed with the HTK software tools [1] and was designed to recognize all 39 phones. Results with monophone and biphone systems are reported.

2. BACKGROUND

The main frame-based features used in this work are Discrete Cosine Transform Coefficients (DCTCs). These coefficients were obtained by encoding the global spectral shape of the frame using a DCT. The process can be summarized as follow.

First, a frame of speech was emphasized by a second-order banpass IIR filter whose center frequency was at 3200 Hz. The filtered signal was then windowed by a Kaiser window function prior to FFT analysis. A 512-point FFT was used and the logarithmic-scaled spectral amplitude was calculated. A set of DCT basis vectors (13 in this paper) was computed to encode the spectral shape over a selected frequency range. A dot product between the spectrum and the basis vectors yielded the DCTCs for the frame.

To compute spectral/temporal features, a Discrete Cosine Series was used to represent the temporal characteristics of the speech over several consecutive frames. A detailed description for computing DCTC and DCSC features can be found in [2].

3. METHODS

This section discusses the three proposed methods of non-uniform time sampling for spectral/temporal feature computation. Each method assumes that the frame-based or static parameters have already been obtained. Thus all methods are discussed from the viewpoint of including temporal information into the final feature set.

3.1. Variable Block Length Method

Use of a DCT to capture the temporal information is a method for encoding the signal trajectory over a period of time with a small number of parameters. Moreover, encoded information can be reconstructed, but generally with error. This method could thus be used a method for lossy encoding. In the work presented here, the method uses reconstruction distortion as the criteria to determine block length.

Beginning with an error threshold, the appropriate block length is determined with an iterative process. First the shortest block length within the range of consideration is encoded and then decoded. If the reconstruction error is less than the selected error threshold, then a larger block is tried. The largest block size that yields an error below the threshold is selected. Note that the error was normalized to reflect average error per frame rather than total error over the length of the block.

The effect of this algorithm is to use shorter blocks in regions of rapid spectral change and longer blocks in regions of more steady spectra. For example, a short block length is used in rapid transitions, since these regions have a high reconstruction error.

Figure 1 shows an example of this method applied to real speech data. An utterance of the letter "b" is displayed in the spectrogram. As can be observed from the figure, large blocks and short blocks are used in different areas. In particular, the shortest blocks are in the beginning and end regions, with much longer blocks for the steady-state vowel portion of the utterance.

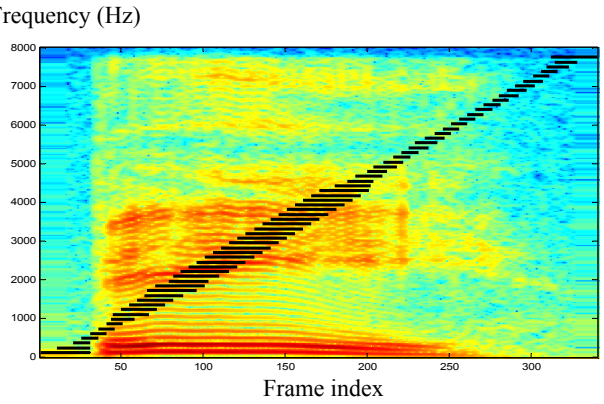

Figure 1: Illustration of the variable block length technique for the letter "b."

3.2. Variable Block Spacing Method

This method emphasizes or de-emphasizes an area by means of applying different block spacings. A short block spacing is selected in areas where the spectrum changes rapidly and long block spacings in regions with more steady spectra. The spectral transition measurement was performed using a method presented by Furui [3], beginning with the DCTCs for each frame-based observations. The space between the two adjacent blocks was determined by examination of the amplitude of the spectral derivative associated with the interval in the block. If the spectral derivative is high, which implies abrupt transitions, a short block spacing is used and if the spectral derivative is low, a long spacing is used.

Figure 2 shows an example of this method applied to an utterance of the letter "b." Blocks are closely spaced in the beginning and end regions where there are rapid spectral changes, but blocks are spaced much father apart in the steady vowel portion. The effect is a variation of that shown in Figure 1.

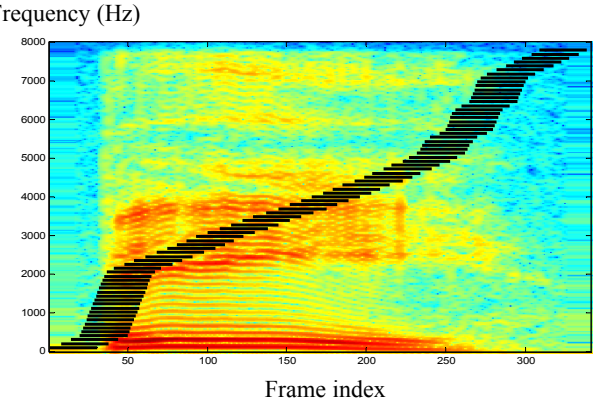

Figure 2: Illustration of the variable block spacing technique for the letter "b."

3.3. Combined Block Length Method

This method encodes the temporal information using a short and a long block at the same time. Encoded parameters for the two blocks are combined and then Principal Components Analysis (PCA) is used to select the most important features. Thus, for this method, PCA selects the features which are most important: short trajectory features, long trajectory features, or a mixture of both.

4. PHONE RECOGNITION TASK AND DATABASE

All methods were tested with phone recognition using the TIMIT database. The NIST training set, comprised of 3696 sx and si sentences, was used for training data and the NIST core test set, which consists of 192 sx and si sentences, was used for testing. Only test results are reported.

In this paper, all 61 TIMIT phones were folded into a 48 phone set as described by Lee in [4]. For the case of monophone models, there were 48 HMM models, one each per phone. However result evaluations were performed with 39 phones in which confusions among some phones were not counted as errors. This arrangement is also described in [4].

The monophones were initialized with time-labeled data which was provided with the TIMIT database using 20 iterations of the Viterbi and 20 iterations of the B-W algorithms. The initialized models were further trained in the embedded mode with the B-W algorithm for 5 iterations. Tests were made with both full covariance and diagonal covariance matrix HMMs. Note that the HTK was used to construct the HMM based recognizers.

A simple right-context biphone models was also built to test all methods. The occurrences of all phone pairs in the training data were counted and the top 400 most frequently occurring pairs were selected. In addition, the 48 monophones were added resulting in a total of 448 models. Each biphone model was cloned from its corresponding well-trained monophone. Five iterations of the B-W algorithm were applied to train the biphone models.

Each HMM used in this paper was a 3-state left-to-right model with mixtures of Guassian densities. Only self-transitions and transitions to the next state were allowed. Bigram language modeling was applied at the phone level.

5. EXPERIMENTS

5.1. Static Features

All experiments reported in this section (except ones with MFCC parameters) used spectral/temporal featured that were extracted from frame-based parameters as follows. An analysis frame size of 25 ms with a frame spacing of 2 ms was used. The frequency range was selected as 70 Hz to 7000 Hz for each spectrum. Thirteen DCTC parameters were extracted from each frame.

5.2. Control Experiments

As a control experiment, the fixed block length and fixed block spacing method was used. Each frame feature was expanded by 3 Discrete Cosine Series over time resulting in a total of 39 terms per block.

An experiment was also conducted to evaluate the recognizer with MFCC features. Parameters used to compute MFCCs were mostly the default values of the HTK toolkit. This can be summarized as follow. The frame size was 25 ms and the frame spacing was 10 ms. For each frame, 12 MFCCs were computed from 20 band pass filters spread from 70 Hz to 7000 Hz. A normalized energy term was also computed. The final 39 terms were obtained by appending the delta terms and the delta-delta terms to the original MFCCs and energy term.

For each method proposed, a series of experiments was performed. Every method (except the MFCC as just mentioned) used 3 DCT basis vectors over time to encode the temporal information within each block. We kept the number of features the same for each method so that we could fairly compare the results from various methods.

To find reasonable sets of parameters for each method, experiments were first conducted with simple monophone models. For these tests, each HMM had 3 mixtures per state and a diagonal covariance matrix.

Once a reasonably good set of parameters had been found, we then tested the methods with more complex models including models with more mixtures per state, models with a full covariance matrix and context-dependent models.

Figure 3 depicts best recognition accuracies achieved by each method using monophone systems with diagonal covariance matrices and full covariance matrices.

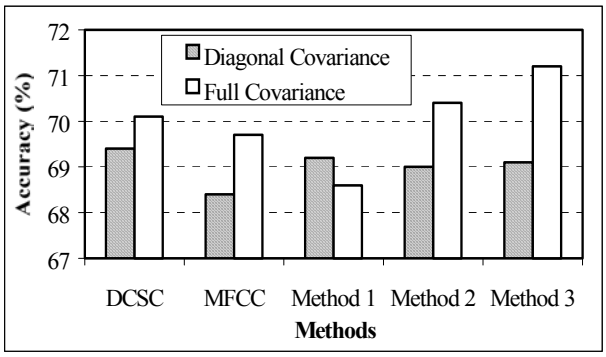

Figure 3: Monophone results of different methods with diagonal and full covariance mixtures.

From Figure 3, the results labeled "DCSC" were computed with uniform time spacing. The results labeled "MFCC" were also obtained with uniform time spacing, using the MFFC terms and

associated dynamic terms as mentioned above. Method 1 represents the variable block length method; Method 2 represents the variable block spacing method; and Method 3 represents the combined block length method. In most cases, use of a full covariance matrix gave higher accuracy than use of diagonal covariance matrix. With the diagonal covariance systems no methods were better than the DCSC method. However, Method 3 was best with a full covariance matrix.

Best results for each method with context-dependent models are shown in Table 1. Note that full covariance mixtures were not used in context-dependent experiments because the models would have been too complex and we could not find a reasonable way to reliably train all models. Thus experiments were conducted using diagonal covariance matrices only. It can be seen from the table that only Method 3 yielded higher results than the standard DCSC method.

Method	Accuracy (%)
DCSC	71.5
MFCC	69.1
Variable Block. Length	71.6
Variable Block Spacing	71.5
Combined Block Length	72.6

Table 1: Results of different methods with right-context biphone models

5.3. Comparison of Results

The absolute best result on the NIST core test set was 74.4% as reported in [5]. Table 2 summarizes the methods used and accuracy achieved by other systems.

Interestingly, our monophone system with a full covariance matrix was nearly as good as most systems. The accuracy obtained from such a system was 71.2%.

System	Accuracy on NIST core test set (%)
Robinson, 1994 [6]	73.9
Deng and Sameti, 1996 [7]	73.4
Mari et al., 1996, [8]	68.8
Chang and Glass, 1997, [9]	73.4
Ming and Smith, 1998, [5]	74.4
Karnjanadecha and Zahorian	72.6

Table 2: Phonetic recognition accuracies on NIST core test set for various ASR systems.

5. CONCLUSIONS

The variable block length and the variable block spacing methods presented in this paper did not result in any improvements over uniform time spacing methods. The use of two block lengths, followed by a PCA transformation, did perform slightly better than the baseline system. Thus, the experimental results do not support our original hypothesis that simple schemes for non-uniform time spacings for speech feature computations will more closely reflect information content in speech and thus improve HMM-based recognizers.

However, to our knowledge, the result obtained with our monophone system was the highest among the results reported in the literature. The context-independent models yielded higher results than most context-dependent system in the past. This system should be straightforward to duplicate, since a "standard" HMM and simple context-dependent models were used. We believe that recognition performance of our front end could be improved with a more refined HMM and with carefully designed context-dependent models. It is also possible that non-uniform time spacing methods can be more carefully integrated into the HMM framework and would result in more advantages.

6. REFERENCES

[1] Young, S. J., Odell, J., Ollason, D., Valtchev, V., and Woodland, P., Hidden Markov Model Toolkit V2.1 reference manual, Technical report, Speech group, Cambridge University Engineering Department, March 1997.

[2] Zahorian S. A., and Nossair Z. B., "A Partitioned Neural Network Approach for Vowel Classification Using Smoothed Time/Frequency Features," *IEEE Trans. Speech and Audio Processing*, vol. 7, pp. 414-425, July 1999.

[3] Furui, S., "On the role of spectral transition for speech perception," *J. Acoust. Soc. Am.*, 80, pp. 1016-1025, October 1986.

[4] Lee, K. F., "Large Vocabulary Speaker-Independent Continuous Speech Recognition: The SPHINX System," Ph.D. dissertation, Computer Science Department, Carnegie Mellon University, 1988.

[5] Ming, J. and Smith, F., "Improved Phone Recognition Using Baysian Triphone Models," *Proc. ICASSP 98*, pp. 409-412, Seattle, WA, May 1998.

[6] Robinson, A., "An Application of Recurrent Nets to Phone Probability Estimation," *IRRR Trans. Neural Networks*, vol. 5, pp. 298-305, March 1994.

[7] Deng, L. and Sameti, H., "Transitional Speech Units and their Representation by Regress Markov States: Application to Speech Recognition," *IEEE Trans. SAP*, vol. 4, pp. 301-306, 1996.

[8] Mari, J., Fohr, D. and Junqua, J., "A Second-Order HMM for High Performance Word and Phoneme-Based Continuous Speech Recognition," *Proc. ICASSP 96*, pp.435-438, 1996.

[9] Chang, J. W. and glass, J., "Segmentation and Modeling in Segment-based Recognition," *Proc. Eurospeech 97*, pp. 1199-1202, Rhodes, Greece, September 1997.

7. ACKNOWLEDGMENTS

This work was partially supported by NSF grant BES-9977260.