MULTIPLE TIME RESOLUTION ANALYSIS OF SPEECH SIGNAL USING MCE TRAINING WITH APPLICATION TO SPEECH RECOGNITION

Spiros Dimopoulos[1], *Alexandros Potamianos*[1], *Eric-Fosler Lussier*[2] *and Chin-Hui Lee*[3]

[1]Dept. of Electronics and Computer Engineering, Technical University of Crete, Chania, Greece
[2]Department of Computer Science and Engineering, The Ohio State University, Columbus, OH
[3]School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA
{sdim, potam}@telecom.tuc.gr, fosler@cse.ohio-state.edu, chl@ece.gatech.edu

ABSTRACT

In this paper, we propose two methods of multiple time-resolution analysis of speech and their application to Automatic Speech Recognition (ASR). Constant frame-rate multi-scale analysis is proposed based on a box of multi-scale features. Then a variable rate analysis is proposed based on the selection of the optimal temporal resolution on the fly by a properly trained non-linear classifier unit. The classifier's parameters are trained using the discriminative method of Minimum Classification Error (MCE) training. We use the recently proposed Conditional Random Fields (CRF) phonetic recognition system that effectively combines highly correlated features. Results are reported on a frame-wise classification task and also on TIMIT phone recognition task. Results show that (i) CRFs can effectively combine multi-scale features and (ii) MCE trained variable rate CRFs are competitive with the "box" combination method.

Index Terms— ASR, MCE, Conditional Random Fields, Variable Frame Rate, Multiple Frame Rates

1. INTRODUCTION

The theoretical concept that supports the variable temporal resolution analysis is that speech is a dynamic phenomenon producing signals that are sometimes stationary to durations that can reach 100 msec during some vowels, but can also have drastically changing spectrum in the scale of a couple of milliseconds during stop consonants and phoneme transition segments. Also some phonemes, especially reduced phonemes, have very short durations. Additionally, the speaker's rate of speech can affect the rate at which the speech signal is changing [1]. The analysis of speech signal according to the rate of speech and subsequent recognition system adaptation is a step towards knowledge rich speech modeling [2]. The trade-off between stationary segments (low temporal variation - higher spectral resolution) and fast changing segments (high temporal variation - lower spectral resolution) analysis leads to constant temporal resolution analysis typically of 10 msec frame rate and 25 msec window size. Our goal is to extend this common speech front-end processing into either a multi-resolution or a variable-resolution speech front-end. The first one incorporates speech features from different resolutions into a constant frame-rate feature vector. Because the feature vector's rate is still constant, it does not require any further modifications to the available speech recognition engines. The second and the most promising is the variable-resolution front-end and a subsequent variable-rate speech recognition system. We no longer process the incoming speech signal at a constant frame-rate/window size framework, but we vary the analysis according to the rate of change of speech.

Different approaches have been employed in order to achieve this multiple resolution in temporal processing. In [3], multi-stream Hidden Markov Models (HMM) were used to incorporate different rate features into the recognition system. In [1], acoustic models were trained at different rates and then a N-Best list rescoring with a phone-dependent posterior-like score was used to determine the appropriate model rate for a segment of speech. In [4], words with multiple dictionary entries for each rate of speech were used, composed by sub-word units trained at different rates. Finally in [5], frame-rate selection was used by frame picking using Mel-Frequency Cepstrum Coefficient (MFCC) feature distance.

In this work, we keep the main concepts of the variable rate analysis, but also equip our system with a Rate of Speech (ROS) metric that is computed by a non-linear-combination of the rate of change of sub-spectral regions and a classifier that selects the optimal frame-rate / window size (FR/WS), taking into consideration the ROS Metric. For the training of the parameters we use the Minimum Classification Error (MCE) / Generalized Probabilistic Descent (GPD) method [6][7][8]. The method iteratively updates the model parameters, in order to minimize the classification error rate on a cross-validation set. For the speech recognition experiments, we use the recently proposed Conditional Random Fields (CRF) approach to phonetic recognition [9].

The paper is organized as follows: In Section 2, first a simple multi-rate system is presented and then our approach to a variable temporal resolution speech recognition system is analytically presented. In Section 3 is described implementation of the system. In Section 4, the experiments and the results are presented and we conclude in Section 5.

2. MULTIPLE TEMPORAL SCALES PROCESSING

2.1. Multiple Time Resolution Analysis of Speech

First multiple frame rate / window size speech features were combined and used in a speech recognition task. The multiple time-scale analysis was achieved by including different temporal analysis features in a constant frame-rate box. The resulting feature vector was calculated at constant frame-rate, but the features inside the box were computed at different frame-rates. Using this method, we appended the standard single Frame Rate / Window Size method with a number of signal parameters from other resolutions. The resulting feature vector is highly correlated, so a de-correlation step was deemed necessary. Different combinations of features were tested in order to find the appropriate one. Results on this method are presented in Section 4.

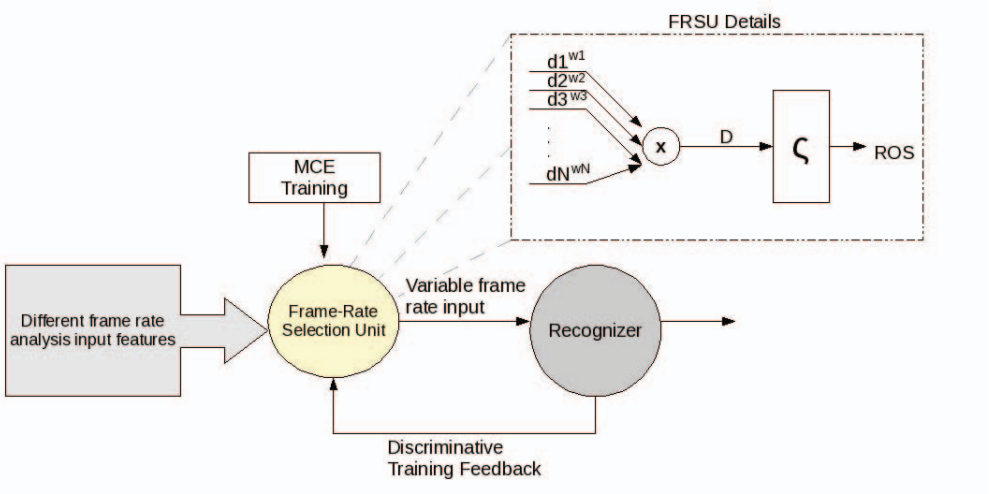

Fig. 1. Variable frame rate speech recognition system

2.2. Variable Time Resolution Analysis of Speech

We propose a functional system that implements the variable frame rate analysis in Automatic Speech Recognition (ASR). The system works in a closed-loop in order to select the best frame rate for each temporal segment of speech. The general concept is shown in Fig. 1. The main unit of interest is the Frame-Rate Selection Unit (FRSU) which is trained to select the best frame rate for a specific segment of speech. The training of the FRSU's parameters is done using an MCE discriminative training method. The training is done in two passes as explained in Sections 2.3 and 2.4.

2.3. Spectral Change Metric

We compute a metric for each subspectral region of a segment of speech signal that indicates the rate of change for that spectral region. By using spectral regions we include information about the rate of change of different regions. The combination of these metrics forms a global spectral distance metric that models the rate of change of speech. The rate of change can be computed from a Fast Fourier Transformation (FFT) analysis and then taking the first-order time difference for each region. The combination of these partial spectral change metrics benefits from a non-linear combination method such as the product rule with weights that can be trained with a Maximum Likelihood (ML) or MCE method. Non-linear method can detect changes in minor spectral regions which indicate a transition segment compared to a rather smoothed result of a linear combination method. The global spectral distance is computed by the equation:

$$D = \prod d_i^{w_i} \tag{1}$$

with w_i the trainable weights and d_i the distance for the i-th spectral region. The weighting parameters can be trained in a first-pass MCE training of the FRSU unit.

2.4. Mapping to Rate Of Speech

Next we want to map the computed distance metric to the optimal FR/WS pair. The mapping function of choice is a sigmoid with a few parameters that can be trained in a second-pass MCE training phase:

$$ROS = a + \frac{c}{1 + e^{-b(D+d)}} \tag{2}$$

with ROS meaning the Rate Of Speech, D the global spectral distance computed above and a,b,c and d trainable parameters learning the non-linear mapping from the spectral change distance metric to the optimal rate of speech. After these two processing steps, we have a continuous function of the rate of change of speech. This can be used to select the appropriate FR/WS pair for each speech segment.

2.5. Training of parameters using an MCE method

Now that we have described the transfer function of the FRSU unit, we can proceed in describing the learning process that we use to train the free parameters on Eq. (1) and (2). An obvious solution to the learning process is the ML estimation of the parameters. We include this estimation method as the baseline in our work. Given the ML estimated model parameters, we use the MCE training method to improve the parameter estimates.

The task of MCE training can be divided into three main steps:

a) Choose a discriminant function for the description of the classification task.

b) Create a misclassification measure to express the classifier decision process.

c) Form a cost function that would be an indicator of the classification's success.

All the above quantities must be continuous and differentiable with respect to the estimated parameters. The classifier can be designed to have a simple discriminant function of the form:

$$g_j(X;\lambda) = |ROS_j - ROS_X| \tag{3}$$

with ROS_j indicating the j-th Rate Of Speech prototype value, and ROS_X the Rate Of Speech Metric computed as shown in Sections 2.3 and 2.4. With the previous formulation of the discriminant function, we can state the classifier's decision process as:

$$C(X) = C_i, \text{ if } g_i(X;\lambda) = min_j(g_j(X;\lambda)) \tag{4}$$

The next step in MCE formulation is the definition of a class misclassification measure, which in fact expresses the decision rule

in Eq. (4) in a functional form [6]. We choose the rather frequently used measure:

$$d_j(X;\lambda) = g_j(X;\lambda) - \left[\frac{1}{M-1}\sum_{k,k\neq j} g_k(X;\lambda)^{\eta}\right]^{\frac{1}{\eta}} \quad (5)$$

with η a positive smoothing constant and M the number of classes. When the misclassification measure is way below zero, this indicates a correct classification. Instead when it is positive, it indicates an incorrect classification.

After we have defined the misclassification measure, we create the cost/loss function. The function must be continuous and indicative of the classification's success/error rate. We choose the sigmoid mapping function which is bound between 0 and 1:

$$l_j(X;\lambda) = \frac{1}{1+exp(-\gamma d_j(X;\lambda))}, \gamma > 1 \quad (6)$$

with γ the sigmoid's scaling factor. This loss function is a smooth and continuous measure of the classification task's success. When sample X is correctly classified then the misclassification measure decreases way below 0 and the loss function approaches 0. When it is incorrectly classified the misclassification measure indicates the level of failure and the loss function approaches 1. Now we can evaluate the classifier's performance on an unknown sample X using the following smooth function:

$$l(X;\lambda) = \sum_{i=1}^{M} l_i(X;\lambda)1(X \in C_i) \quad (7)$$

where $1()$ is the indicator function and is 1 when sample X belongs to class i else is 0.

The next concern is a minimization method for the expected loss of the classifier during training, in order to estimate the appropriate values for the free parameters on Eqs. (1) and (2). We want to minimize the expected loss which is:

$$L(\lambda) = E_X\{l(X;\lambda)\} = \sum_{i=1}^{M}\int_{X\in C_i} l_i(X;\lambda)p(X)dX \quad (8)$$

with X summing over all samples of a training set. We use the GPD algorithm with parameter space transformations in order to impose constrains on the free parameters [6]. In practice we minimize the empirical loss assigning equal probability mass to each sample. The empirical loss will converge to the expected loss if a training set of sufficient size is used. The general update equation of the parameter set we are training (λ) at a given iteration of the process (t) is:

$$\lambda_{t+1} = \lambda_t - \varepsilon\nabla l(X;\lambda)|_{\lambda=\lambda_t} \quad (9)$$

with ϵ the learning coefficient. We can use a 2-pass training procedure. In the 1st pass, keep the parameters of Eq. (2) constant and train the paramaters of Eq. (1). In the 2nd pass use the values found during 1st pass and train the parameters on Eq. (2). As an example of MCE/GPD iterative update, the update equation for parameter b on Eq. (2), when sample $X \in C_i$, is given:

$$b_{t+1} = b_t - \varepsilon\frac{\partial l_i(X;\lambda)}{\partial b} \quad (10)$$

with

$$\frac{\partial l_i(X;\lambda)}{\partial b} = \frac{\partial l_i}{\partial d_i}\cdot\frac{\partial d_i}{\partial g_i}\cdot\frac{\partial g_i}{\partial ROS_X}\cdot\frac{\partial ROS_X}{\partial b} \quad (11)$$

A similar partial derivative chain rule can be used in order to derive the update equations of the other parameters of the FRSU module.

3. IMPLEMENTATION

First we computed the spectral change metric. The quantity used to compute the spectral region differences was the Mel-Scale Spectral Magnitude, computed by 20 channel filterbank analysis. We run a first pass recognition task using multiple FR/WS pairs on a cross validation set. We segmented each utterance to 30 msec segments and labeled each segment using the frame classification results of the first pass. Every segment that was classified correctly under one FR/WS pair, it was labeled with its corresponding ROS label (ROS=10 for FR/WS=10/25msec, etc.). Then an ML training was done using the distance metric as input and the ROS label as output. The parameters of Eq. (1) and Eq. (2) were trained using ML estimation. We keep this mapping as a baseline.

Next we implemented an MCE/GPD embedded iterative training algorithm and done a re-training of the mapping functions. In the first pass, we re-train the parameters of Eq. (1). Lower frequencies are mapped with larger weights and higher frequencies tend to have smaller weights. In the second pass we re-train the parameters of Eq. (2). To simplify the training procedure we kept the parameter a to value 5 and parameter c to value 5 in order to have a uniform dynamic range of frame rates, i.e, we re-train parameters b and d. We use a maximum of 30 iterations.

The mapping function that emerged, as the result of this iterative procedure, was used to select the frame-rate on a frame-based classification task and also the normal recognition task. Frame classification and utterance recognition results on ML and MCE trained FRSUs were then compared.

4. EXPERIMENTS

For our experiments we used the TIMIT speech database. Although the addition of noise seems to favor variable rate systems [5], we work on clean data. We used a training set, a cross-validation set, a core test set and an extended (core+rest) test set as described in [9]. For the classification and recognition tasks, we used the HMM and the CRF frameworks.

For HMMs we used the HTK Toolbox and trained 48 Context-Indepented (CI) 3-state 16-mixture monophone models on the training set and performed the multiple rate recognition task on the other sets. For comparison we performed the same recognition task on CRF using 48 CI 1-state monophone models. Then a reduction to 39 phonemes on both frameworks was done for comparison.

The features we are using are MFCC with delta and acceleration (MFCC_D_A) combined from different FR/WS pairs, (fr,ws)={(10,25), (5,12.5), (2.5,6.25)} in msec, as described in subsection 2.1, shown as MFCC-MFR in the following tables. We also report results on MFCC_D_A computed at 10 msec as baseline, shown as MFCC-10.

Table 1: Multi-Rate *Phoneme Recognition* Results Box Combination Method

Method	Feature group	Core Set		Ext Set	
		Acc %	Rec %	Acc %	Rec %
HMM	MFCC-10	49.08	52.77	48.98	53.15
HMM	MFCC-MFR	48.35	53.55	49.03	54.50
CRF	MFCC-10	47.33	51.87	47.22	51.78
CRF	MFCC-MFR	**50.94**	**58.11**	**51.00**	**58.49**

From Table 1, we can see that HMMs have better performance when using plain single-rate MFCC_D_A parameters. When using multi-scale MFCC_D_A features, HMMs cannot integrate effeciently this extra information. In contrast, the CRFs improve performance by combining efficiently these highly correlated parameters.

Next we proceed using the CRF framework and combine 44 phonological class posteriors computed from Multi-Layer Perceptrons (MLP) as described in [9] with single-rate MFCC_D_A and also with multirate MFCC_D_A. We also include results using exclusively phonological posteriors as a baseline.

Table 2: Multi-Rate *Phoneme Recognition* Results Using Phonological Posteriors - CRF Framework - Box Method

Feature group	Core Set		Ext Set	
	Acc %	Rec %	Acc %	Rec %
Posteriors	66.72	68.68	68.16	70.32
Posteriors+MFCC-10	68.25	71.12	**69.86**	72.86
Posteriors+MFCC-MFR	**68.74**	**72.83**	**69.86**	**74.20**

The results in Table 2, show the CRFs ability to integrate features of different quality and time-scale. When Posteriors are merged with single-rate MFCCs, an improvement is clear. Adding multi-rate MFCCs improves somewhat the results. The improvement in recognition is significant.

Next we present the results using the CRF framework for the variable rate system. We used the classification results from the cross-validation set to train the FRSU with the MCE method as described earlier in Section 2. We performed a variable rate experiment on the Core and Ext sets using the ML trained FRSU as baseline and also the MCE trained FRSU, in order to select the optimal FR/WS for each 30msec segment. In this experiment, we used MFCC_D_A features combined from different rates and windows (fr,ws)={(10,25),(7.5,18.75),(5,12.5)} in msec. We report the results for a *frame-based classification* task. We also have included frame-level results from the baseline single-rate MFCC system and the multi-rate Box method using MFCC and Posteriors+MFCC. We have to note here that only MFCC features were computed at multiple rates, Posteriors were interpolated versions of the constant rate version. Also the TIMIT *phoneme recognition* task results are presented in Table 3.

Table 3: Variable & Multi-Rate Speech Recognition System *Frame Classification* Results - CRF Framework

System setup	Features	Core %	Ext %
Const-Rate	MFCC-10	50.06	48.93
Box Method	MFCC-MFR	54.54	53.85
	Post+MFCC-MFR	70.07	70.70
ML FRSU	MFCC	49.81	49.36
	Post+MFCC	61.37	61.94
MCE FRSU	MFCC	51.70	51.10
	Post+MFCC	**72.40**	**73.11**

Table 4: Variable Speech Recognition System *Phoneme Recognition Results* - CRF Framework

System	Features	Core Set		Ext Set	
		Acc %	Rec %	Acc %	Rec %
ML	MFCC	47.88	51.47	48.76	51.35
FRSU	Post+MFCC	62.50	65.61	64.57	68.15
MCE	MFCC	46.48	51.14	46.75	51.81
FRSU	Post+MFCC	**67.49**	**70.33**	**68.81**	**72.41**

From the results in Table 3 for the *frame-based classification* task, we see that the MCE method outperforms the ML method. Also the box method performs well using multi-rate MFCC features. Finally the MCE method is the best performer when combining Posteriors and MFCCs at different rates. Looking at the *phoneme recognition* results in Table 4, we see the ML method is performing better that the MCE method when using MFCCs. MCE performs better when using Posteriors and MFCCs. Also comparing Table 2 and Table 4, we see that the box method performs better than the best variable rate method (MCE) during recognition. This indicates again the CRFs ability to take the most out of highly correlated features and also the variable rate system's weakness of interpreting good frame classification results to equally good phoneme recognition results.

5. CONCLUSIONS AND FUTURE WORK

This work is the first step towards incorporating multiple temporal resolution processing features in the speech recognition system. First we merged features from different time-scales into a static FR/WS system as part of a multi-rate feature vector. The results show small but significant improvement compared to the single frame-rate system. The CRF framework integrates multi-rate highly correlated features better than HMMs. Then we proposed a functional Variable FR/WS system that uses a ROS Metric from the non-linear combination of sub-spectral speech regions change rate and a FRSU to select the optimal frame-rate locally for a segment of speech. We trained our system using ML and MCE methods. The MCE method outperforms the ML and the box method in the frame classification task. In normal recognition task the box method gave slightly better results than the MCE method. Two reasons are mainly to blame: (i) the large number of insertions and (ii) the mapping from frames to state and phoneme sequences.

We are working towards minimizing the error rate of recognition task, by controlling the insertions of the variable rate system and the translation of variable rate frames to states and phonemes. A variable rate system based on phonological posterior features computed at different temporal resolutions should be implemented for improved results.

6. REFERENCES

[1] Venkata R. Gadde, Kemal Sonmez, and Horacio Franco, "Multirate asr models for phone-class dependent n-best list rescoring," ASRU, 2005.

[2] Chin-Hui Lee, "From knowledge-ignorant to knowledge-rich modeling: a new speech research paradigm for next generation automatic speech recognition," ICSLP, 2002.

[3] Astrid Hagen and Herve Bourlard, "Using multiple time scales in the framework of multi-stream speech recognition system," Eurospeech, 1997.

[4] Jing Zheng, Horacio Franco, and Andreas Stolcke, "Rate of speech for large vocabulary conversational speech recognition," ASR2000.

[5] Qifeng Zhu and Abeer Alwan, "On the use of variable frame rate analysis in speech recognition," ICASSP, 2000.

[6] Biing-Hwang Juang, Wu Chou, and Chin-Hui Lee, "Minimum classification error rate methods for speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 3, pp. 257–265, May 1997.

[7] Biing-Hwang Juang and Shigeru Katagiri, "Discriminative learning for minimum error classification," *IEEE Transactions on Signal Processing*, vol. 40, pp. 3043–3054, December 1992.

[8] Shigeru Katagiri, Biing-Hwang Juang, and Chin-Hui Lee, "Pattern recognition using a family of design algorithms based upon the generalized probabilistic descent method.," *Proceedings of The IEEE*, vol. 86, no. 11, pp. 2345–2373, November 1998.

[9] Jeremy Morris and Eric Fosler-Lussier, "Conditional random fields for integrating local discriminative classifiers," *IEEE Transactions on Acoustics, Speech, and Language Processing*, vol. 16, no. 3, pp. 617–628, March 2008.