# Speaking Rate Dependent Multiple Acoustic Models Using Continuous Frame Rate Normalization

Sung Min Ban and Hyung Soon Kim
Pusan National University, Busan, Korea
E-mail: {bansungmin, kimhs}@pusan.ac.kr Tel: +82-51-510-1704

***Abstract*—This paper proposes a method using speaking rate dependent multiple acoustic models for speech recognition. In this method, multiple acoustic models with various speaking rates are generated. Among them, the optimal acoustic model relevant to the speaking rate of test data is selected and used in recognition. To simulate the various speaking rates for the multiple acoustic models, we use the variable frame shift size considering the speaking rate of each utterance instead of applying a flat frame shift size to all training utterances. The continuous frame rate normalization (CFRN) is applied to each of training utterances to control the frame shift size. Experimental results show that the proposed method outperforms both the baseline and the conventional CFRN on test utterances.**

## I. Introduction

As speech recognizer is deployed to a variety of people and environments, the problems that should be overcome have been emerged. These problems include the variabilities due to additive noise, channel distortion and speaker characteristics, and speaking rate variability should also be considered. Previous studies revealed that the difference between speaking rates of training and test data degrades the performance of speech recognition [1], but the researches on this problem are relatively limited.

Some researches have been performed to relieve the performance degradation caused by speaking rate variability. A kind of adaptation technique to adjust the state transition probability was proposed to deal with this problem [1], but most studies were performed in terms of variable frame rate in feature domain [2]-[5]. In early researches on the variable frame rate approach, frame shift size was controlled by using frame dropping. Whether a certain frame is dropped or not is decided by speaking rate of the utterance. To measure the speaking rate, various techniques have been proposed [2]-[7]. Among them, Euclidean distance between two feature vectors of adjacent frames was used to measure the speaking rate [2]. To consider the information from multiple frames, the norm of the cepstral derivative or entropy of cepstra was also employed for the estimation of speaking rate [3], [4]. In addition, frame dropping was performed using short frame shift size such as 2.5 ms in order to capture the dynamic changes in high speaking rate region [5]. As an alternative approach, phonetic information from the speech recognizer was used for the estimation of speaking rate. In this approach, there was a method of considering the phonetic classes in the test utterance [6], and recently, the average phone durations of training and test utterances were used to normalize the frame shift size of the test data to the training data [7].

Because the speaking rate is variable according to speaker and task, always there exists the mismatch between the speaking rates of training and test data. Most conventional algorithms focused on the speaking rate adaption of test data to the training data. But it is difficult to obtain the representative frame shift size of training data, because the variance of the average phone duration in all training utterances is large. Consequently, it is not very effective to adapt the speaking rate of test data to that of training data.

In this study, instead of adapting the speaking rate of test data to that of training data, multiple acoustic models are generated from the feature sets with various frame shift sizes, and the acoustic model which is appropriate to the test data is selected and used in recognition. If a flat frame shift size is applied to all of the utterances, the average phone duration is changed, but its distribution is unchanged. Thus, as mentioned earlier, the problem of large variance in average phone duration is still remained. In this paper, to alleviate this problem, the continuous frame rate normalization (CFRN) is applied to all of the training data in order to reduce the variance of the average phone duration.

This paper is organized as follows: In section 2, speaking rate dependent multiple acoustic models is introduced and CFRN is described in section 3. Finally, the performance of the proposed algorithm is evaluated in section 4, and the conclusion of this paper is drawn.

## II. Speaking rate dependent multiple acoustic models

To adapt the speaking rate of test data to that of the training data, CFRN was proposed recently. In previous study of CFRN, average phone duration of test utterance was used as the speaking rate information. But the average phone duration in training utterances has a broad distribution as shown in Fig. 1. Therefore, even if the speaking rate of test data is matched to the global average phone duration of training data with an optimal frame shift size, the normalized speaking rates of each phoneme can be so different.

To overcome this problem, this paper proposes a method using speaking rate dependent multiple acoustic models. The block diagram in Fig. 2 shows the proposed method. After obtaining the speaking rate of each training utterance, CFRN

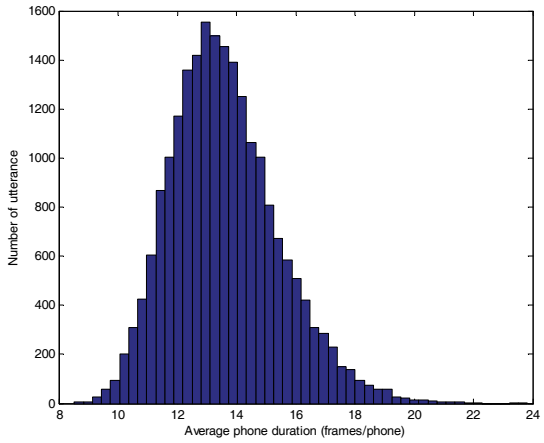


Fig. 1. *The distribution of the average phone duration in training utterances.*

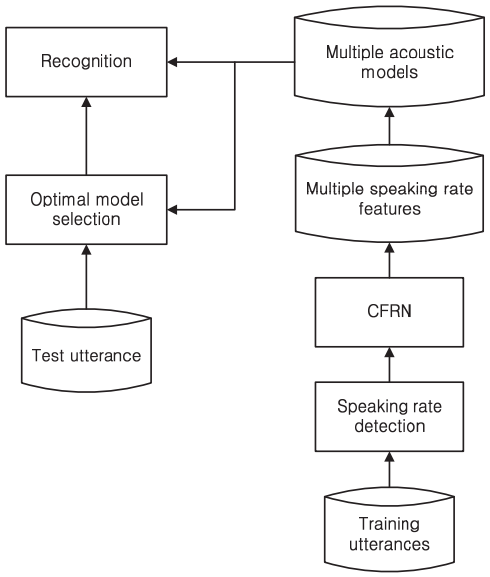


Fig. 2. *Block diagram of the proposed method.*

is applied to each training utterance. Then the speaking rate normalized feature vectors corresponding to the target speaking rate are generated by using an appropriately chosen frame shift rate. In this way, multiple acoustic models corresponding to each of frame shift sizes are generated and is used in recognition. Given the test utterance, an optimal acoustic model relevant to this test data is selected based on the maximum likelihood criterion. In Fig. 2, to estimate the speaking rate of the training utterance, the average phone duration obtained by the forced alignment is used. The average phone duration $f(i)$ of the $i$-th utterance is defined as

$$f(i) = \sum_{j=1}^{M_i} t_i(j)/M_i. \tag{1}$$

Here, $M_i$ is the number of phones in the $i$-th utterance, $t_i(j)$ represents the number of frames at the $j$-th phone in the $i$-th utterance. In next session CFRN will be explained in some detail.

## III. CFRN FOR ACOUSTIC MODEL TRAINING

At first, the conventional CFRN is described. In the conventional CFRN, to normalize the speaking rate of the test utterance to that of training data, target speaking rate from training utterances is obtained as

$$\Phi = \sum_{i=1}^{N}\sum_{j=1}^{M_i} t_i(j) / \sum_{i=1}^{N} M_i \tag{2}$$

where, $\Phi$ is the global average phone duration, $i$ is the utterance index, and $N$ is the number of total utterances in training data [7]. Using the $\Phi$ and average phone duration $\widehat{f}_{test}(i)$ from the $i$-th test utterance, warping factor $warp(i)$ is obtained as follows:

$$warp(i) = \begin{cases} min_{warp} & if \ \ \frac{\widehat{f}_{test}(i)}{\Phi} \leq min_{warp} \\ max_{warp} & if \ \ \frac{\widehat{f}_{test}(i)}{\Phi} \geq max_{warp} \\ \frac{\widehat{f}_{test}(i)}{\Phi} & otherwise \end{cases} \tag{3}$$

where $warp(i)$ is the warping factor of the $i$-th test utterance to normalize speaking rate of the $i$-th test utterance to the original speaking rate. The minimum and maximum warping factor are set as $min_{warp}$ and $max_{warp}$ to avoid the warping factor to be unrealistic value. Then the frame shift size of the $i$-th utterance is expressed as

$$s(i) = warp(i)T_{step} \tag{4}$$

where $T_{step}$ is the original frame shift size.

In our proposed method, the CFRN is applied to the training utterances and the warping factor in the equation (3) is modified to $warp_{w_k}(i)$ as

$$warp_{w_k}(i) = \begin{cases} min_{warp} & if \ \ w_k\frac{\widehat{f}_{train}(i)}{\Phi} \leq min_{warp} \\ max_{warp} & if \ \ w_k\frac{\widehat{f}_{train}(i)}{\Phi} \geq max_{warp} \\ w_k\frac{\widehat{f}_{train}(i)}{\Phi} & otherwise \end{cases} \tag{5}$$

where $\widehat{f}_{train}(i)$ represents the average phone duration of the $i$-th training utterance. $warp_{w_k}(i)$ is the warping factor to normalize the speaking rate of the $i$-th training utterance to the speaking rate corresponding to the $k$-th target ratio $w_k$. To generate multiple acoustic models, $w_k$ has multiple values. If the average phone duration of a certain training utterance is large, then its warping factor is increased more than that of other utterance having relatively small average phone duration. Then frame shift size $s_{w_k}(i)$ of the $i$-th utterance can be expressed as

$$s_{w_k}(i) = warp_{w_k}(i)T_{step}. \tag{6}$$

In the case of flat frame shift size, $warp_{w_k}(i)$ becomes a fixed value of $w_k$ regardless of the speaking rate of input training utterance,

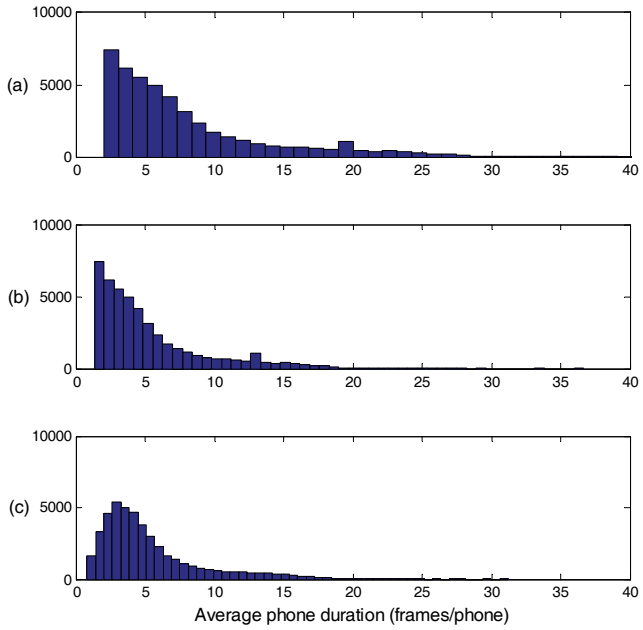


Fig. 3. *Distribution of the duration in phoneme [o] (a) with frame shift rate of 10 ms (b) with flat frame shift rate (c) with variable frame shift rate using CFRN*

The optimal acoustic model to the test data is selected as

$$\widehat{\Psi} = \underset{\Psi_k}{argmax}\, p(\mathbf{x}|\Psi_k). \tag{7}$$

Here, $\Psi_k$ is the one of the multiple acoustic models corresponding to the $k$-th speaking rate. $\widehat{\Psi}$ is the optimal acoustic model relevant to the speaking rate of the test utterance. $p(\mathbf{x}|\Psi_k)$ represents the likelihood of the acoustic model to $\mathbf{x}$, the feature vector sequence of the test utterance.

When CFRN is applied to the training data using $w_k$ as 1.5, the distribution of the duration in phoneme [o] is shown in Fig. 3. In the case of applying the flat shift rate to the training data as in [6], the mean of the phone duration has moved to the other value, but the characteristics of the distribution is unchanged. That is, the normalized data does not have the phone durations concentrated around the target rate but has diverse phone durations. In the case of applying CFRN to the training data, it is observed that the variance of the phone duration is reduced. This means that the normalized data is concentrated around the target speaking rate. Table 1 indicates the standard deviations of the durations in the various phonemes according to the two different normalization techniques. From the table, it is also observed that the normalized data by applying CFRN to the training data is more concentrated around the target speaking rate than that with flat shift rate.

## IV. Experimental results

To evaluate the proposed algorithm, we performed isolated word recognition test. As training data, the phonetically balanced sentence database (PBS DB) is used, which contains 21246 sentences (30 hours) from 100 males and 100 females.

TABLE I
*Standard deviations of the durations (frames) in some phonemes*

| phoneme | [a] | [e] | [i] | [o] | [u] |
|---|---|---|---|---|---|
| flat | 18.81 | 5.55 | 9.59 | 14.56 | 4.39 |
| CFRN | 18.55 | 5.28 | 8.58 | 13.58 | 4.03 |

As test data, the phonetically balanced word database (PBW DB) is used, which is composed of 31640 words (18 hours) from 38 males and 32 females [8] and contains 452 vocabulary words per each speaker. Both database are Korean speech database provided by Speech Information Technology and Industry Promotion Center (SiTEC), Korea. The proposed algorithm is compared with the conventional CFRN method. In this evaluation, 5 multiple acoustic models with 5 frame shift rates are used. In acoustic model training, two different frame shift sizes are included in each of the multiple training sets and also 10 ms frame shift size is commonly included in all the multiple training sets to have stable acoustic model performance in speech recognition. Therefore, each of the multiple training sets includes much more data than baseline model. In equation (3), we set $min_{warp}$ to 0.5 and $max_{warp}$ to 1.5. For training, 3 state left-to-right HMM is used to model triphone based speech unit. Baseline acoustic model uses only 10 ms frame shift size containing 2086 tied states, and each of multiple acoustic models contain about 2609 tied states. We use 14 Gaussian mixtures for each tied state. For feature vector, we use 39-dimensional vector composed of 13-dimensional MFCCs, their delta and delta-delta coefficients, including normalized energy.

Fig. 4 represents the word accuracy according to the average phone duration range. It is observed that the word accuracy tends to decrease as the average phone duration increases. This tendency is due to the fact that as the speaking rate of test word is fast, it becomes closer to the speaking rate of training sentence which is generally more faster than isolated word. Overall, the proposed method outperforms the conventional CFRN and baseline results. Table 2 shows the average word accuracy and error rate reduction. In this table, "proposed(flat)" and "proposed(var)" represent the proposed speaking rate dependent multiple acoustic models applying a flat frame shift size to all training utterances, and those applying the variable frame shift size to each utterance according to the speaking rate, respectively. Both of the two proposed methods outperform the conventional CFRN. From the table, the performance improvements of frame rate dependent algorithms over baseline results are statistically significant, e.g., p-value$<$0.05 for CFRN, p-value$<$0.01 for 'Proposed(flat)', p-value$<$0.001 for 'Proposed(var)'. But contrary to our expectation, applying the variable frame shift size to each utterance is not sufficiently effective. This result seems due to that the proposed normalization method performed in sentence level does not guarantee the normalization of speaking rate in phoneme level. In fact, adjusting only the frame shift size is not a sufficient tool to deal with the performance degradation due to the speaking rate variability, which requires more elaborate

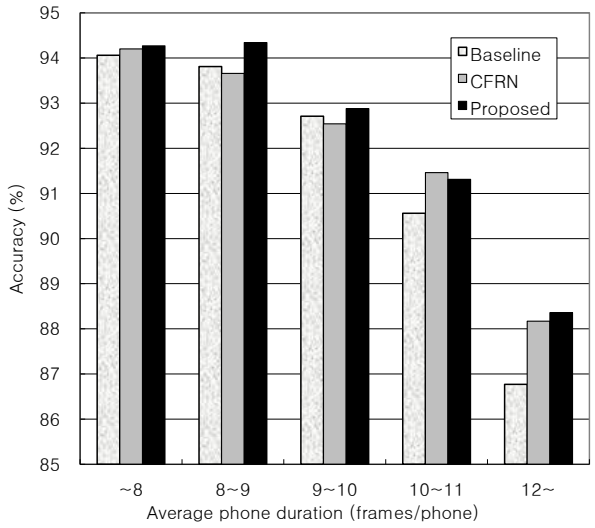


Fig. 4. *Word accuracy according to the average phone duration*

schemes such as phoneme-class dependent duration control and preservation of transient information.

TABLE II
*Average word accuracy and error rate reduction*

| Algorithm | Average accuracy(%) | ERR(%) |
|---|---|---|
| Baseline | 91.41 | - |
| CFRN | 91.88 | 5.30 |
| Proposed(flat) | 92.06 | 7.34 |
| Proposed(var) | 92.13 | 8.19 |

## V. Conclusions

This paper proposed the speaking rate dependent multiple acoustic models to overcome the performance degradation of speech recognition due to the difference between speaking rates of training and test data. Proposed method applied the CFRN technique to the training utterances to make multiple acoustic models with various frame shift sizes. In isolated word recognition task, it was shown that proposed method outperformed both the baseline and the conventional CFRN. As a future work, we are going to normalize the speaking rate in more detailed level instead of using sentence level speaking rate only, and to combine our method with other approach such as an adaptation technique considering the speaking rate [9]. Additionally, we are going to simplify the model selection algorithm, though it is not critical in the condition that parallel processing is available.

## Acknowledgment

This work was supported by the ETRI R&D Program of KCC(Korea Communications Commission), Korea [11921-03001, ”Development of Beyond Smart TV Technology”].

## References

[1] M. A. Siegler, and R. M. Stern, “On the effects of speech rate in large vocabulary speech recognition systems,” in Proc. ICASSP, pp. 612-615, May 1995.
[2] S. M. Peeling and K. M. Ponting, “Variable frame rate analysis in the ARM continuous speech recognition system,” Speech Comm., vol. 10, pp. 155-162, 1991.
[3] P. L. Cerf and D. V. Compernolle, “A new variable frame rate analysis method for speech recognition,” IEEE Signal Processing Letter, vol. 1, no. 12, pp. 185-187, Dec. 1994.
[4] H. You, Q. Zhu, and A. Alwan, “Entropy-based variable frame rate analysis of speech signals and its application to ASR,” in Proc. ICASSP, pp. 549-552, May 2004.
[5] Q. Zhu and A. Alwan, “On the use of variable frame rate analysis in speech recognition,” in Proc. ICASSP, pp. 1783-1786, June 2000.
[6] V. R. Gadde, K. Sonmez, and H. Franco, “Multirate ASR models for phone-class dependent N-best list rescoring,” in Proc. ASRU, pp. 157-161, Nov. 2005.
[7] S. M. Chu and K. Povey, “Speaking rate adaptation using continuous frame rate normalization,” in Proc. ICASSP, pp. 4306-4309, Mar. 2010.
[8] Y.-J Lee, B.-W. Kim, J.-J Kim, O.-Y. Yang, and S.-Y. Lim, “Some considerations for construction of PBW set,” in Proc. of the 12th Workshop on Speech Communications and Signal Processing. Acoustical Society of Korea, pp. 310-314, June 1995.
[9] H. Nanjo and T. Kawahara, “Speaking-rate dependent decoding and adaptation for spontaneous lecture speech recognition,” in Proc. ICASSP, pp. 725-728, May 2002.