Speaking Rate Normalization with Lattice-based Context-dependent Phoneme Duration Modeling for Personalized Speech Recognizers on Mobile Devices

Ching-Feng Yeh[1], *Hung-Yi Lee*[2] *and Lin-Shan Lee*[1]

[1] Graduate Institute of Communication Engineering, National Taiwan University, Taiwan
[2] Research Center for Information Technology Innovation, Academia Sinica, Taiwan
andrew.yeh.1987@gmail.com, tlkagkb93901106@gmail.com

Abstract

Voice access of cloud applications including social networks using mobile devices becomes attractive today. And personalized speech recognizers over mobile devices become feasible because most mobile devices have only a single user. Speaking rate variation is known to be an important source of performance degradation for spontaneous speech recognition. Speaking rate is speaker dependent, it changes from time to time for every speaker. Furthermore, the speaking rate variation pattern is unique for each speaker. An approach of continuous frame rate normalization (CFRN) [1] was recently proposed to take care of the speaking rate variation problem. In this paper, we further proposed an extended version of CFRN for personalized speech recognizers on mobile platforms. In this approach, we use context-dependent phoneme duration models adapted to each speaker to estimate the speaking rate utterance by utterance based on lattices obtained with a first-pass recognizer. The proposed approach was evaluated on both read speech and spontaneous recordings from mobile platforms and significant improvement were observed in the experimental result.

Index Terms: speaking rate, mobile, speech recognition

1. Introduction

With advances and popularity of mobile devices such as cell phones and tablet computers, people nowadays spend much more time with these devices than the past, for which voice is definitely the most convenient and desired interface. Physical keyboard is too large while virtual keyboard on the screen is too small. Voice interaction can be fast, easy and natural for users from children to elders, with the computation requirements achievable by most existing mobile devices.

When voice access of cloud applications via mobile devices [2] is considered, the relatively mature technologies such as voice command or keyword spotting are no longer adequate, but spontaneous continuous speech recognition with large vocabulary is essential. Because most devices are used by a single user, a highly personalized speech recognizer can be developed for each user with his own mobile devices, which may include highly personalized components such as personalized acoustic models, personalized language model and personalized lexicon [3]. In other words, mobile application scenario can be an ideal platform for speech recognizers.

For speech recognition, in addition to the speaker variability, noise disturbances and problems with spontaneous speech, an extra issue which was less investigated is the speaking rate variation [4]. Speaking rate is not only speaker dependent, but changes significantly from time to time for the same speaker, especially for spontaneous speech produced for mobile applications. In fact, each speaker has his own speaking rate variation patterns. It has been well known for long that the mismatch in speaking rate between training and testing data may seriously degrade the recognition performance. Many approaches were proposed to handle this issue, some focused on signal processing [5][6], some on ensemble modeling [7][8][9][10], etc. Continuous frame rate normalization (CFRN) [1] was recently proposed to normalize the frame rate of feature extraction.

This paper presents an extension of CFRN with lattice-based context-dependent phoneme duration modeling adapted to each speaker. The proposed approach was evaluated on corpora of read speech and spontaneous mobile recordings. Significant improvement was observed in the evaluation.

The rest of this paper is organized as follows. The corpora for experiments are introduced in section 2, and a brief review of CFRN is given in section 3. The proposed lattice-based context-dependent phoneme duration modeling is described in section 4. Experimental setup, results and analysis are reported in section 5 with final conclusion presented in section 6.

2. Corpora for Experiments

The corpora used for the experiments reported here consist of two collections of recordings, as listed in Table 1. The first one (ASTMIC) is a Mandarin read speech corpus consisting of utterances produced by 200 gender-balanced speakers, with a total length of 31.8 hours. The content of the utterances are news sentences. We divide this corpus into two subsets, one for constructing speaker-independent acoustic models and the other for testing, respectively labeled as "Train" and "Test" in Table 1. The other corpus is PlurkRec, which is a collection of spontaneous Mandarin speech recorded by various types of mobile phones. The content of the utterances are personal updates the speakers tried to post to Plurk, a popular social network website. The topic distribution within the corpus is highly diverse, with many different types of background noise and varying channel distortion due to mismatched microphones of different mobile phones. These characteristics are close to the scenario for real-world mobile applications.

There are 170 utterances available for each speaker in both ASTMIC and PlurkRec test sets. For each speaker, we randomly pick 50 utterances for adaptation, another 50 for development and the rest 70 for evaluation.

Table 1. *Details of Corpora for Experiments*

Corpus	**ASTMIC**		**PlurkRec**
Set	Train	Test	Test
Speech Type	*Read*	*Read*	*Spontaneous*
# Speakers	190	10	15
# Utterances	32185	1700	2550

3. Previously Proposed CFRN

Continuous frame rate normalization (CFRN) [1] was recently proposed to handle the speaking rate mismatch through adjusting the window size and window shift during feature extraction. The speaking rate is defined based on the utterance-level average phoneme duration, which is the total occupied duration (number of frames) of non-silence speech divided by the total number of phonemes in the utterance,

$$D_{avg}(u_i) = \frac{\sum_{j=1}^{J_i} D(p_j)}{J_i}, \quad (1)$$

where u_i is the i-th utterance, J_i is the total number of non-silence phonemes in u_i, p_j is a phoneme within u_i, and $D(p_j)$ is the duration of p_j.

Then the average phoneme duration Φ over the training data is calculated similarly,

$$\Phi = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J_i} D(p_j)}{\sum_{i=1}^{I} J_i}, \quad (2)$$

where I is the total number of utterances in the training data.

The speaking rate $R(u_i)$ for an utterance u_i can then be computed as in (3),

$$R(u_i) = \frac{\Phi}{D_{avg}(u_i)}, \quad (3)$$

so an average phoneme duration shorter than the training set average leads to a speaking rate larger than 1.0. The feature vectors for the utterances are then extracted with warped window size and shift based on the upper and lower bounded speaking rate in (4)(5)(6), which implies shorter average phone duration leads to higher speaking rate.

Once we have the speaking rate for each utterance, the features for the utterances are extracted with warped window size and shift. The warping factor is defined in (4):

$$R'(u_i) = \begin{cases} R_{max} & R(u_i) > R_{max} \\ R_{min} & R(u_i) < R_{min} \\ R(u_i) & otherwise \end{cases}, \quad (4)$$

where R_{max} , R_{min} are upper/lower bounds of the speaking rate,

$$w_{CFRN} = w_o \,/\, R'(u_i), \quad (5)$$

$$s_{CFRN} = s_o \,/\, R'(u_i), \quad (6)$$

where w_o and s_o are the original window size and shift, typically 25 ms and 10 ms.

4. Proposed Approach

The proposed warped feature extraction is illustrated in Fig. 1. For each input speech utterance at the upper-right corner of Fig. 1, feature extraction is first performed with the original normal window size and shift. A first-pass recognition is then applied to generate the lattice to be used for speaking rate estimation as shown in the middle-left of Fig. 1. Personalized context-dependent phoneme durations adapted to the user are then used in the lattice-based warping factor computation, with output used for the warped feature extraction for second-pass recognition.

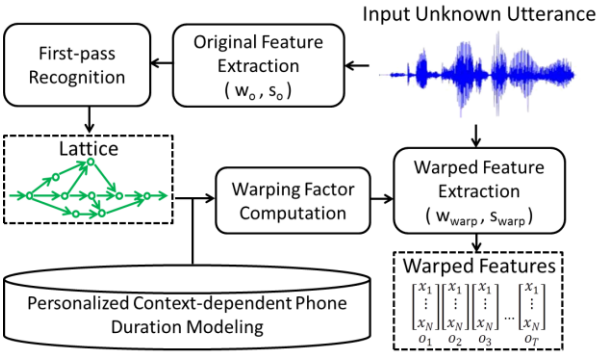

Figure 1: *Proposed Warped Feature Extraction Framework*

4.1. Personalized context-dependent phoneme duration modeling

Context-dependent phoneme duration modeling has been popularly used in speech synthesis [11][12], which is used here. By aligning feature vectors and acoustic models for the utterances in the training set, we can have durations for each phoneme in the transcriptions, with which context-dependent phoneme (triphone) duration model can be trained by decision tree clustering as illustrated in Fig. 2. For the scenario considered in this work, there is no sufficient data to build speaker-dependent duration models for each speaker. Therefore, we used phoneme duration data from multiple speakers to train initial speaker-independent duration models and then adopted MLLR to generate speaker-adapted duration models for each speaker.

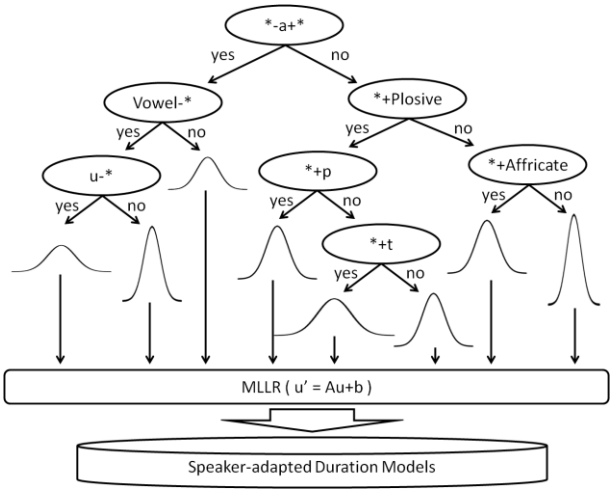

Figure 2: *Context-dependent Phoneme Duration Modeling with Decision Tree Clustering*

In Fig. 2, each node of the decision tree is a question about context information for a phoneme cluster, and decisions for further classification are made in these nodes. By considering factors including entropy and training data availability, duration models can be well-constructed for both seen and unseen triphones. Different from the well-known approach in acoustic modeling, here only a one-dimensional single Gaussian is trained for the duration of each triphone.

It is well-known that different phonemes have different duration distributions. A good example is shown in Fig. 3 for duration histograms of vowels for four different left context phoneme classes: plosives, affricates, voiced consonants and vowels, estimated with the training set of ASTMIC listed in Table 1. It is clear that there exist significant difference between duration distributions for different left context phoneme classes. As a result, it is necessary not only to build models for different phonemes but to model context-dependent phoneme durations as well, since the phoneme duration highly depends on the context.

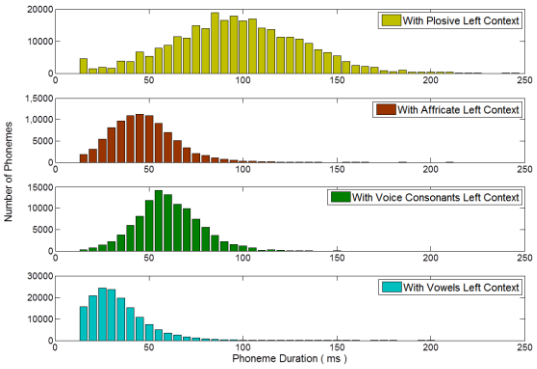

Figure 3: *Duration Histograms for Vowels with Different Context Phoneme Classes in Training Data*

4.2. Lattice-based speaking rate estimation

For an unknown utterance u_i entered, in order to have transcriptions for better estimate of the speaking rate, a first-pass recognition is performed with features extracted using the original window size and shift to produce a lattice. The expected number of frames for each complete path l_i in the lattice can then be obtained as in (7) :

$$D_{exp}(l_i) = \sum_j D(p_{j-1} - p_j + p_{j+1}), \qquad (7)$$

where $\bar{p} = (p_1, p_2 \ldots p_J)$ is the corresponding phoneme sequence for the path l_i in the lattice and $D(p_{j-1} - p_j + p_{j+1})$ is the mean of the personalized context-dependent phoneme model for triphone $p_{j-1} - p_j + p_{j+1}$ for the models obtained in section 4.1.

The expected duration (or number of frames) of the utterance u_i is then estimated as in (8) :

$$D_{exp}(u_i) = \frac{\sum_s P(\bar{o}_i|l_s)P(l_s)D_{exp}(l_s)}{\sum_s P(\bar{o}_i|l_s)P(l_s)}, \qquad (8)$$

where $\bar{o}_i$ represents the feature vector sequence of u_i extracted with the original window size and shift, $D_{exp}(l_s)$ is the expected duration of l_s as in (7), $P(\bar{o}_i|l_s)$ and $P(l_s)$ are the likelihood function and prior probability obtained from the acoustic and language models in the first-pass recognizer. Therefore, the expected duration for u_i in (8) is the average of the expected durations for all possible paths l_s in the lattice weighted by the corresponding posterior probabilities.

The personalized lattice-based speaking rate $\hat{R}(u_i)$ for an utterance u_i is then estimated as the ratio of the expected duration (number of frames) $D_{exp}(u_i)$ of the utterance u_i to the actual duration $D_o(u_i)$ for it as in (9) :

$$\hat{R}(u_i) = \frac{D_{exp}(u_i)}{D_o(u_i)} \ , \qquad (9)$$

where $D_{exp}(u_i)$ and $D_o(u_i)$ are both for the frames extracted with the original window size and shift. With this personalized speaking rate for the utterance u_i, the new window size and shift for u_i are estimated just as (5)(6),

$$\hat{w} = w_o \ / \ \hat{R}(u_i) \quad , \qquad (10)$$

$$\hat{s} = s_o \ / \ \hat{R}(u_i) \quad , \qquad (11)$$

In other words, $\hat{w}$ and $\hat{s}$ should be shorter if $\hat{R}(u_i) > 1.0$ or the speaking rate is higher.

5. Experiment

5.1. Experimental environment setup

The corpora used for experiments were mentioned in section 2. Acoustic features adopted here were 39-dim Mel-scale frequency cepstral coefficients (MFCC) with delta and delta-delta coefficients. In addition, cepstral mean variance normalization (CMVN) was applied to enhance the robustness against noise in the collected mobile recordings.

Due to the huge diversity of topic distribution among the evaluation data, including ASTMIC and PlurkRec, appropriate lexicon and training or adaptation data for language modeling is not yet available at the moment when this paper is submitted. Therefore, only recognition results on syllable level or the syllable error rate (SER) are reported below. Because there exists special mapping relation between syllables and characters in Mandarin, syllable error rate is highly correlated and parallel to character error rate (CER) in Mandarin speech recognition. So such results similarly indicated the achievable recognition performance even though the character error rate is still not available.

Bigram syllable model trained with syllable transcriptions of broadcast news were used as the language model here. The lexicon contains all 404 phonologically allowed Mandarin syllables with the corresponding phoneme sequence.

Standard state-tied triphone structure is used for acoustic model training. For model adaptation, class-based MLLR is applied for each test speaker.

5.2. Experimental results

Experimental results for syllable error rates of read and spontaneous speech are listed in Table 2. In row (1), the error rate was high especially for mobile recordings due to the mismatched recording environments. After MLLR was applied, the error rates dropped evidently for both types of speech in row (2). In row (3), significant improvement can be observed for spontaneous speech after CFRN, which indicated that taking speaking rate distortion into consideration is useful. However, for read speech, the performance was slightly degraded even for tuned parameters with the development set. The results for the proposed method are listed in rows (4) and (5). In row (4), we assume only the 1-best path existed in the lattice, so the only difference from CFRN in row (3) is the context-dependent duration modeling. The results indicated that slight improvement was achieved for spontaneous speech while insignificant degradation was observed for read speech. The results of the complete proposed approach illustrated in Fig. 2 are listed in row (5). Compared with row (4), using lattice instead of 1-best transcription brought significant improvements

for spontaneous speech. The result validated the assumption that recognized 1-best transcription could be biased. Comparing the proposed approach in row (5) with CFRN in row (3), evident improvement was achieved.

Table 2. *Experimental Results (SER) (%)*

	Test (Read)	Test (Spontaneous)
(1) SI	29.10 %	46.30 %
(2) MLLR	21.88 %	34.64 %
(3) CFRN	21.98 %	34.18 %
(4) Duration + 1-best	22.00 %	33.94 %
(5) Duration + Lattice	22.05 %	33.38 %

Fig. 4 shows the comparison among different methods for utterances with different speaking rate. The horizontal axis is the speaking rate $\hat{R}(u_i)$ in (9) for each utterance and the vertical axis represents SER for utterances with speaking rates in the range. In general, the error rate was lowest in the middle and highest at most right for spontaneous speech, which is consistent to our observation on relationship between speaking rate and error rate. The speaking rate normalization approaches, including both CFRN and the proposed method, gave the most improvements for utterances with high distortion as expected. However, for normal speaking rates, these methods did not bring much enhancement and could be risky due to wrong estimation of the warping factor. In such cases, the proposed method did not cause significant degradation while CFRN did sometimes.

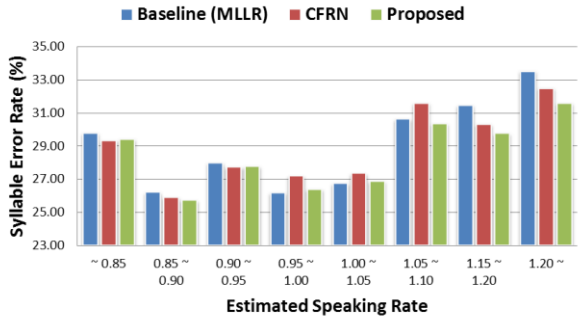

Figure 4: *Syllable Error Rate using Different Methods with Respect to Estimated Speaking Rate*

5.3. Analysis of experimental results

To further analyze the nature of spontaneous speech, 2-D histograms for read speech and spontaneous speech were shown in Fig. 5. The horizontal axis stands for the number of syllables in reference transcriptions and the vertical axis represents the number of syllables in the recognized results. The color shows the count of utterances with the corresponding horizontal and vertical values. It is evident that the results for read speech were more centralized and fitted with the diagonal line. That is, the number of syllables in recognized results and reference transcriptions were the same at most times. However, for spontaneous speech, the results were dispersed and deviated from the diagonal line, primarily below the diagonal line. So the numbers of syllables in recognized results were very often less than those in reference. This could be caused by higher speaking rate, since the numbers of frames were not enough to generate the recognized result with equal number of syllables to the reference transcription. This is just why frame rate normalization helped for spontaneous speech.

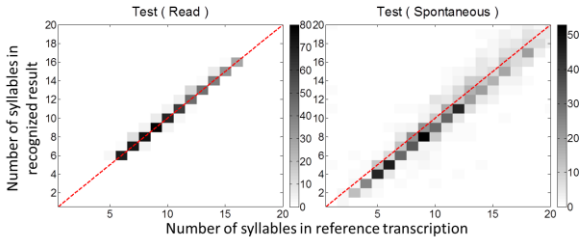

Figure 5: *2-D Histogram of Utterance with Respect to Numbers of Syllables in Results and Reference*

The analysis for phoneme numbers and utterance durations for read and spontaneous speech is shown in Fig. 6. In Fig. 6, each data point represents a single utterance. The horizontal axis stands for the number of non-silence phonemes, and the vertical axis represents the summation of occupied duration by these phonemes. The data points are painted and marked according to the set they belong to. Data points with gray crosses come from ASTMIC training data in read speech. Those with red circles come from ASTMIC testing data in read speech. And those with yellow circles come from PlurkRec testing data in spontaneous speech. Regression lines in corresponding color are also plotted in the figure. It is obvious that the slope of regression line for spontaneous speech is smaller than read speech, which indicates speaking rate is a serious issue on mobile devices. Furthermore, variation of number of phonemes and utterance duration is also larger than read speech, thus makes the task difficult.

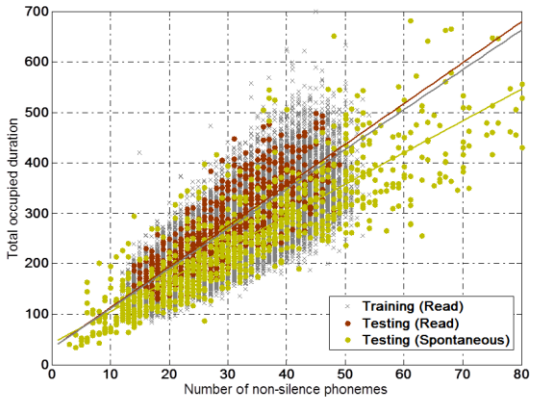

Figure 6: *Scattered Plot of Utterances with Respect to Phoneme Numbers and Occupied Durations*

6. Conclusion

In this paper, we proposed a method utilizing personalized context-dependent phoneme duration modeling over lattices to evaluate the speaking rate and enhance the previously proposed CFRN framework. Through experiments and evaluation, the proposed method was proved effective, especially for speech with high speaking rate distortion. Such approach will be useful for personalized recognizers in mobile application scenario.

7. References

[1] Stephen Chu and Daniel Povey, “Speaking Rate Adaptation using Continuous Frame Rate Normalization”, ICASSP 2010

[2] Dirk Schnelle, “Mobile Speech Recognition”, 2009

[3] Jingjing Liu, Stephanie Seneff, and Victor Zue, “Harvesting and Summarizing User-Generated Content for Advanced Speech-Based Human-Computer Interaction, Special Issue on Advances in Spoken Dialogue Systems and Mobile Interfaces”, Theory and Applications: IEEE Journal of Selected Topics in Signal Processing (JSTSP), 2012.

[4] M. A. Siegler, and R. M. Stern, “On the Effects of Speech Rate in Large Vocabulary Speech Recognition Systems”, ICASSP 1995.

[5] Q. Zhu and A. Alwan, “On the Use of Variable Frame Rate Analysis in Speech Recognition”, ICASSP 2000.

[6] H. You, Q. Zhu, and A. Alwan, “Entropy-based Variable Frame Rate Analysis of Speech Signals and Its Application to ASR”, ICASSP 2004.

[7] H. Nanjo, K. Kato, and T. Kawahara, “Speaking Rate Dependent Acoustic Modeling for Spontaneous Lecture Speech Recognition”, Eurospeech 2001.

[8] K. Okuda, T. Kawahara, and S. Nakamura, “Speaking Rate Compensation Based on Likelihood Criterion in Acoustic Model Training and Decoding”, ICSLP 2002.

[9] Sung Min Ban and Hyung Soon Kim, “Speaking Rate Dependent Multiple Acoustic Models using Continuous Frame Rate Normalization”, APSIPA 2012.

[10] Morgan, Nelson H., “Combining Multiple Estimators of Speaking Rate”, Acoustics, Speech and Signal Processing, Proceedings of the 1998 IEEE International Conference.

[11] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, “Duration Modeling for HMM-based Speech Synthesis”, ICSLP 1998.

[12] Yun Tang; Wenju Liu; Bo Xu "Trigram duration modeling in speech recognition", ISCSLP 2004.