SPEAKING RATE ADAPTATION USING CONTINUOUS FRAME RATE NORMALIZATION

Stephen M. Chu

IBM T. J. Watson Research Center
Yorktown Heights, New York 10598, USA
schu@us.ibm.com

Daniel Povey[1]

Microsoft Research
Redmond, Washington 98052, USA
dpovey@microsoft.com

ABSTRACT

This paper describes a speaking rate adaptation technique for automatic speech recognition. The technique aims to reduce speaking rate variations by applying temporal warping in front-end processing so that the average phone duration in terms of feature frames remains constant. Speaking rate estimates are given by timing information from unadapted decoding outputs. We implement the proposed *continuous frame rate normalization* (CFRN) technique on a state-of-the-art speech recognition architecture, and evaluate it on the most recent GALE broadcast transcription tasks. Results show that CFRN gives consistent improvement on all four separate systems and two different languages. In fact, the reported numbers represent the best decoding error rates of the corresponding test sets. It is further shown that the technique is effective without retraining, and adds little overhead to the multi-pass recognition pipeline found in state-of-the-art transcription systems.

Index Terms – CFRN, speaking rate adaptation, speech recognition, frame rate normalization.

1. INTRODUCTION

An essential problem in *automatic speech recognition* (ASR) is how to adapt to the myriad types of variability in human speech, from low-level acoustic disparities to speaker differences to higher-level linguistic diversities. In fact, most current ASR systems rely on a collection of normalization and adaptation techniques to tackle such variations, e.g., *cepstral mean normalization* (CMN), *vocal tract length normalization* (VTLN), and *maximum likelihood linear regression* (MLLR).

One significant source of variation is the speaking rate, which measures how fast a speech segment is spoken. The overall speaking rate varies at both the speaker level and the utterance level. It may also fluctuate within a sentence. It has been shown that variance in speaking rate has a clear negative impact on speech recognition performance [1].

The *hidden Markov model* (HMM) framework inherently takes care of the problem to some extent by allowing certain degree of freedom in the temporal axis. However, in practice HMMs often model duration inadequately, especially when the speaking rate variance is large. Efforts have been made to address the shortcoming from the modeling perspective, ranging from explicit duration modeling in HMM to more drastic departures from the HMM paradigm such as segment-based recognition [2].

This work aims to address speaking rate variation in the feature space. In particular, we propose a normalization technique for speaking rate that can be used under the standard HMM framework, as an additional tool in conjunction with existing normalization and adaptation procedures in the recognition pipeline.

The proposed technique normalizes different speaking rates by adjusting both the *frame rate* and *frame size* in the acoustic feature extraction stage. Different from prior work in [3]-[7], instead of using a small number of predetermined discrete rates, we allow the frame rate to vary continuously. For speaking rate detection, the work bypasses the low-level signal processing approach [3]-[4], and relies on the ASR system itself for estimation and adaptation. Furthermore, many proposed adaptation schemes require additional training, sometimes of multiple systems [5]-[7]. In contrast, the method presented here is shown to be effective even without matched training. Finally, most results on this subject are reported on small tasks or limited systems. In this paper, we implement the proposed *continuous frame rate normalization* (CFRN) technique in a state-of-the-art ASR architecture, and evaluate it on the most recent GALE broadcast transcription tasks. Results show that CFRN gives consistent improvement on four separate systems for two languages. In fact, these results represent the best error rates reported on the corresponding test sets.

The remainder of this paper is organized as follows. In Section 2, we discuss speaking rate adaptation and introduce CFRN. The system architecture is given in Section 3. Section 4 presents the experimental results, followed by conclusions in Section 5.

2. SPEAKING RATE ADAPTATION WITH CFRN

2.1. Speaking Rate and Frame Rate

Most HMM-based ASR systems use a fixed frame rate and window size in front-end processing. This is based on the assumption that the non-stationary speech signal can be approximated by a piecewise quasi-stationary process. The common choice of 25 ms window and 10 ms step size is a compromise between data rate and resolution determined empirically to give a reasonable performance on average.

Using a fixed frame rate is not optimal on two levels. First, the fixed rate might be too slow to adequately capture the dynam-

[1] The author was at IBM Research while this work was performed.

ics in transient speech events while generating highly correlated redundant observations for stationary speech units [3]. This has motivated efforts to adjust frame rate based on the duration of low-level speech units. In [4], a very short step size of 2.5ms is used to process the speech signal and an entropy measure is used to extract the appropriate frames from the sequence. In [8], signal processing is done at a fixed frame rate, and frames are later dropped or inserted (inferred from existing frames) to normalize the duration of all phones, resulting in an observation sequence with effectively a varying frame rate. In general, using variable frame rates to equalize low-level temporal dynamics often requires dramatic changes from the conventional value.

Second, the overall speed of speaking is variable, and most of the time does not match the optimal operating point of a fixed-frame-rate system. Both the literature and our own experiments clearly indicate that the further the speaking rate drifts, the higher the error rate. It is this overall shift in *speaking rate* that we aim to compensate for in this work. In [5], utterances are grouped into *fast* and *slow* speech, and two different frame rates are chosen. Similarly, three discrete sets of frame rates and window sizes are used in [7]. Fig. 1 shows a typical distribution of speaking rate observed in broadcast speech (consisting of regular broadcast news and the more spontaneous broadcast conversations). Here the speaking rate is defined as the average phone duration in an utterance. The graph confirms that the speaking rate does vary. Further, the speaking rate appears to vary continuously, thus making it difficult to justify any arbitrary thresholds for fast and slow speech. Lastly, the dynamic range of the speaking rate is actually moderate. This suggests that a modest adjustment might be enough to reduce the mismatch between speech rate and frame rate.

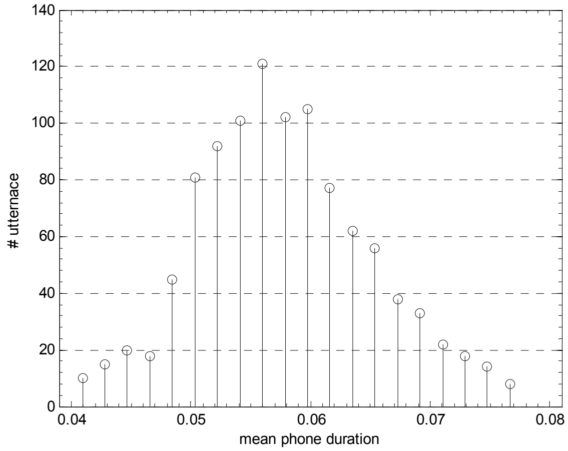

Fig. 1 Distribution of utterances over speaking rates (measured in average phone duration) on GALE *eval*'08 Arabic set.

2.2. Speaking Rate Detection

The reliable detection of speech rate is key to an adaptation scheme. Note that our definition of speaking rate is slightly different from the commonly used *number of syllables per second* (*minute*) in Linguistics. Our objective is to normalize the average speed of pronounced speech, hence, any inter-word silence and other non-speech segments, though important to the perceived rhythm and speed, must be discarded from the calculation.

Our measure of the speaking rate is an average of phone durations in a speech segment. It is therefore important to determine at which level the average should be taken, so that the subsequent adaptation is the most effective. In the context of broadcast speech transcription, the possible levels include per show, per speaker, per utterance, or within utterance. Both the speaker-level and the utterance-level options are implemented. It is found that per utterance adaptation gives consistently higher gains in recognition performance.

Thus, given an utterance i, composed of a sequence of N_i words: $[w_{i,1}, w_{i,2} \ldots w_{i,N_i}]$, the speaking rate $f(i)$ is,

$$f(i) = \sum_{j=1}^{N_i} t(w_{i,j}) \Big/ \sum_{j=1}^{N_i} n(w_{i,j}) \qquad (1)$$

where $t(w_{i,j})$ is the duration of the word $w_{i,j}$ and $n(w_{i,j})$ is the number of phones in the word.

Transcripts and actual word durations are needed to compute $f(i)$. For training, the timing information can be obtained by forced alignment using existing acoustic models. For test utterances, we propose to use the decoding hypotheses from the unadapted system to compute an estimate of the speaking rate, $\hat{f}(i)$.

Given phone-level alignment, the definition given in (1) can be modified to compute *average vowel duration* or *average syllable duration*, both reasonable alternatives for measuring speaking rate. There also exist ways to detect speaking speed directly from the speech using signal processing methods [2]-[4].

One potential drawback of the ASR-based approach is that it requires an additional decoding pass. However, this does not pose a significant problem for most modern ASR systems as using multi-pass decoding to take advantage of adaptation techniques such as MLLR has become standard. In a practical system with *n* decoding passes, decoding output from the *n*-1'th pass can usually be used for speaking rate estimation, with the final pass replaced by CFRN decoding. Thus, the overall number of decoding passes remains unchanged.

2.3. Continuous Frame Rate Normalization

As discussed above, the fixed frame rate in an ASR system is chosen to give the best overall performance for different speaking rates, and is expected to perform well if a specific utterance's measured speaking rate matches the average.

Fig. 2(*a*) shows a scatter of per utterance cumulative phone duration vs. total number of phones, essentially the numerator and denominator terms in (1), over a large Arabic speech set. Clearly, if the speaking rate $f(i)$ were constant, the scatter would reduce to a straight line, and 1/slope is the universal speaking rate. The objective of speaking rate adaptation is to normalize $f(i)$ so that it remains constant for all utterances.

Instead of directly normalizing the speaking rate by manipulating the speech signal, we apply normalization to the frame rate so that the average phone duration in terms of feature frames remains constant.

We define the target speaking rate Φ to be the global average phone duration:

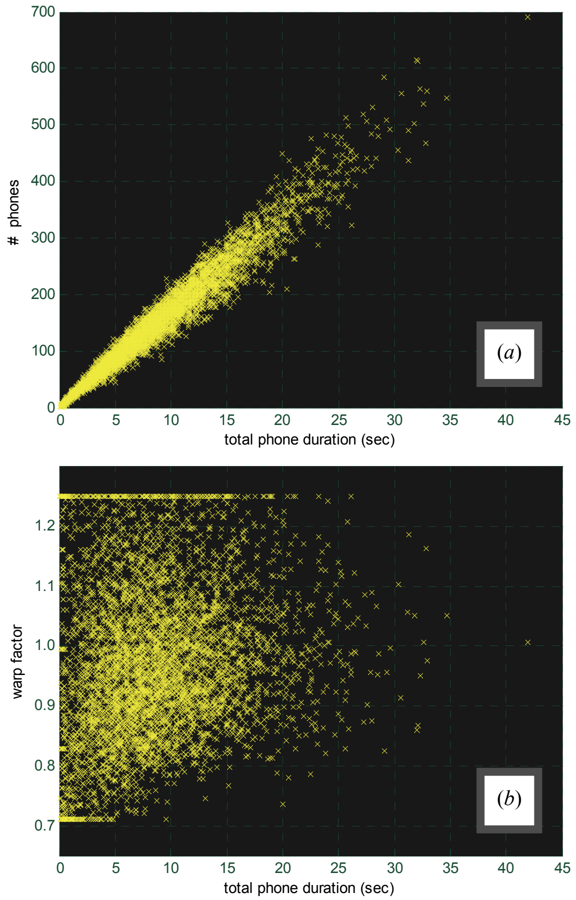

Fig. 2 (a) Scatter plot of number of phones vs. cumulative phone duration per utterance over three GALE Arabic test sets. (*b*) Distribution of the corresponding CFRN warp factors.

$$\Phi = \sum_{i=i}^{M}\sum_{j=1}^{N_i} t(w_{i,j}) \Bigg/ \sum_{i=i}^{M}\sum_{j=1}^{N_i} n(w_{i,j}) \qquad (2)$$

where M is the number of utterances in the entire set. The warping factor for utterance i, $warp(i)$ is the ratio of the speaking rate estimate and the target speaking rate:

$$warp(i) = \begin{cases} \min_{warp}, & \text{if } \dfrac{\hat{f}(i)}{\Phi} \le \min_{warp} \\ \max_{warp}, & \text{if } \dfrac{\hat{f}(i)}{\Phi} \ge \max_{warp} \\ \dfrac{\hat{f}(i)}{\Phi}, & \text{otherwise} \end{cases} \qquad (3)$$

Lower and upper limits $\min_{warp}$ and $\max_{warp}$ are necessary to prevent improbable warping factors caused by unstable speaking rate estimates. Finally, the frame rate is normalized by warping the step size and the window size in the front-end:

$$t_{step}(i) = warp(i) \cdot T_{step},\ t_{win}(i) = warp(i) \cdot T_{win} \qquad (4)$$

where T_{step} and T_{win} are the original fixed step and window sizes, and $t_{step}(i)$ and $t_{win}(i)$ are adapted values for utterance i.

3. SYSTEM ARCHITECTURE

The proposed CFRN technique is implemented on a state-of-the-art multi-pass speech recognition architecture for broadcast transcription. Details of the architecture can be found in [10]. Here we only give a brief summary.

The input audio is sampled at 16 KHz, and coded using 13-dememsional PLP features; 25 ms windows and 10 ms shift are used before CFRN. Nine consecutive frames are spliced and projected to 40 dimensions using LDA and MLLT. The acoustic models are continuous mixture density HMMs with context-dependent states conditioned on cross-word quinphone context. The acoustic models are refined through multiple training stages. The final ML speaker adapted system contains VTLN, fMLLR, and MLLR adaptations. The discriminative system is built with both feature and model space training using either the MPE or MMI criterion. A diagram of the baseline decoding pipeline is shown in Fig. 3.

4. EXPERIMENTS

4.1. Experimental Setup

Experiments are carried out on four setups: *a*. Mandarin speaker independent (SI) system, *b*. Mandarin speaker adapted (SA) system, *c*. unvowelized Arabic system, and *d*. vowelized Arabic system.

Mandarin systems
Acoustic models are built on 1.7K hours of data released by LDC for the GALE program. The SI system has 10K quinphone states modeled by 300K Gaussian densities. The SA system has 15K states and 800K Gaussians. The decoding LM is built by interpolating 20 back-off 4-gram models using modified Kneser-Ney smoothing, and has a107K vocabulary.

Arabic systems
Acoustic models are built on 1.5K hours of transcribed GALE data

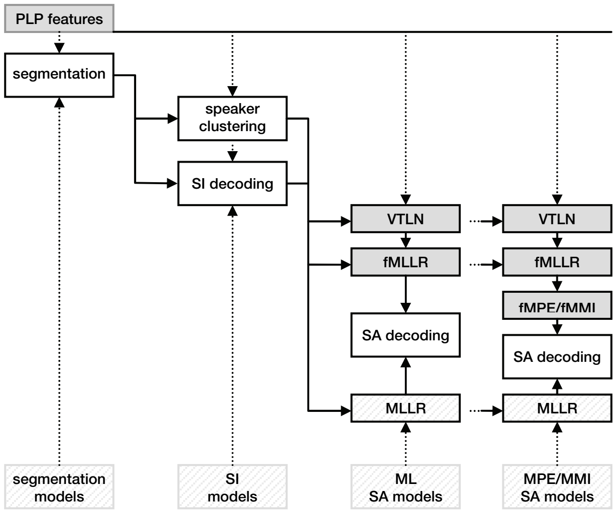

Fig. 3. Baseline decoding pipeline of the IBM broadcast speech transcription system.

from LDC and 85 hours of FBIS and TDT-4 audio. The unvowelized system uses straightforward graphemic models, whereas the vowelized system uses phonetic models and requires inferring short vowels and diacritics missing in written Arabic. The unvowelized system has 5K quinphone states modeled by 400K Gaussian densities. The vowelized system has 6K states and 400K Gaussians. Both systems share an interpolated and smoothed 4-gram LM with a 774K-word vocabulary.

4.2. Experimental Results

Table 1 shows the SI decoding results on the Mandarin dev'08 set defined by the GALE consortium. Three approaches are compared: *a*. the baseline with fixed frame rate, *b*. variable frame rate with three discrete warping factors, and *c*. CFRN. The results confirm that speaking rate adaptation through frame rate adjustment is indeed viable, and continuous warping is more effective, achieving 0.6% absolute reduction in *character error rate* (CER).

Table 1. SI decoding results on GALE mandarin *dev'08*.

	fixed FR	*discrete FR*	*cfrn*
CER	16.9%	16.6%	16.3%
substitutions	13.1%	12.9%	12.6%
deletions	2.6%	2.5%	2.5%
insertions	1.2%	1.2%	1.2%

Two more sets are considered in the full SA decoding tests, shown in Table 2. P-2.5 and eval'08 consist of the un-sequestered data from the GALE phase 2.5 and phase 3 evaluations, respectively. Speaking rates are estimated using the baseline output, and CFRN decoding is carried out subsequently. The results suggest that the gain from CFRN is consistent.

Table 2. SA decoding results on three GALE mandarin test sets.

	dev'08	*p-2.5*	*eval'08*
baseline	7.9%	7.5%	10.5%
+*cfrn*	7.7%	7.4%	10.2%

Table 3 summarizes results on the two Arabic systems. Similar to the Mandarin case, dev'08 is a consortium-defined test set, while eval'07 and eval'08 contain un-sequestered portions of the corresponding GALE evaluation sets. For baseline decoding, the unvowelized system is cross-adapted (with fMLLR/MLLR) on the SA decoding output of the vowelized system, and vice versa. For CFRN decoding, the SA output from the vowelized system is used for speaking rate estimation in both cases. The results show that CFRN is able to consistently give an additional gain on top of the best baselines. Finally, a plot of WER as a function of speaking rate on eval'08 is given in Fig. 4. Corresponding baseline and CFRN WERs are compared. Two observations can be made from the plot. First, off-central speaking rates indeed lead to higher WERs. Second, more consistent gain is made on fast speech.

Table 3. Decoding results on GALE Arabic test sets.

	dev'08	*eval'07*	*eval'08*
unvowelized	13.3%	14.7%	10.9%
+*cfrn*	13.2%	14.6%	10.7%
vowelized	13.3%	14.7%	10.8%
+*cfrn*	13.2%	14.4%	10.6%

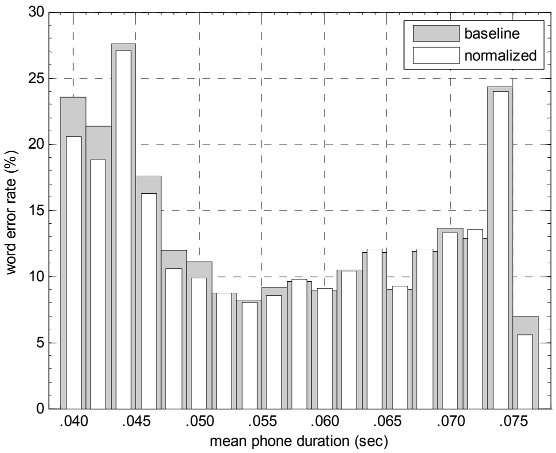

Fig. 4. Error distribution over speaking rate on eval'08 (Arabic). Higher error reduction is observed on fast speech.

5. CONCLUSIONS

This work considers frame rate normalization for speaking rate adaptation in ASR. Applied at test time without retraining, the proposed CFRN technique is shown to consistently reduce error rates over optimized baselines with minimal overhead. In future work, we will investigate nonlinear frame rate warping and matched CFRN training.

REFERENCES

[1] M. A. Siegler, and R. M. Stern, "On the effects of speech rate in large vocabulary speech recognition systems," in *Proc ICASSP'95*, pp. 612-615, May 1995.

[2] J. R. Glass, "A probabilistic framework for segment-based speech recognition," *Computer Speech and Language*, vol. 17, pp. 137-152, 2003.

[3] Q. Zhu and A. Alwan, "On the use of variable frame rate analysis in speech recognition," in *Proc. IEEE ICASSP'00*, pp. 1783-1786, June 2000.

[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'04*, pp. 549-552, May 2004.

[5] H. Nanjo, K. Kato, and T. Kawahara, "Speaking rate dependent acoustic modeling for spontaneous lecture speech recognition," in *Proc. EUROSPEECH'01*, pp.2531--2534, 2001.

[6] K. Okuda, T. Kawahara, and S. Nakamura, "Speaking rate compensation based on likelihood criterion in acoustic model training and decoding," in *Proc. ICSLP'02*, pp. 2589--2592, 2002.

[7] V. R. Gadde, K. Sonmez, and H. Franco, "Multirate ASR models for phone-class dependent N-best list rescoring," in *Proc. ASRU'05*, pp.157-161, November 2005.

[8] J. P. Nedel and R. M. Stern, "Duration normalization for improved recognition of spontaneous and read speech via missing feature methods," in *Proc. ICASSP'01*, pp. 313-316, May 2001.

[9] S. M. Chu *et al.*, "Recent advances in the IBM GALE Mandarin transcription system," in *Proc. ICASSP'08*, pp. 4329-4332, March 2008.