Speaking Rate Estimation Based on Deep Neural Networks

Natalia Tomashenko[1,2] and Yuri Khokhlov[1]

[1] Speech Technology Center, Saint-Petersburg, Russia
www.speechpro.ru
[2] ITMO University, Saint-Petersburg, Russia
{tomashenko-n,khokhlov}@speechpro.com

Abstract. In this paper we propose a method for estimating speaking rate by means of Deep Neural Networks (DNN). The proposed approach is used for speaking rate adaptation of an automatic speech recognition system. The adaptation is performed by changing step in front-end feature processing according to the estimations of speaking rate. Experiments show that adaptation results using the proposed DNN-based speaking rate estimator are better than the results of adaptation using the speaking rate estimator based on the recognition results.

Keywords: speaking rate, speaking rate adaptation, speaking rate estimation, speech recognition, ASR, variable step, DNN.

1 Introduction

Speaking rate is one of significant sources of speech variability. In the past it was shown [1]-[12] that variations in speaking rate can degrade recognition performance in automatic speech recognition (ASR) systems. This effect is typically more significant for fast speech than for slow [1][5], that is, the higher the speaking rate, the higher the error rate in ASR systems. Fast speech is different from normal or slow speech in several aspects [5], such as acoustic-phonetic and phonological characteristics.

Many methods for compensating the effects of speaking rate variability were proposed in the literature [1][5][6][8][9][10]. They include: modification of HMM transition probabilities [6] and modeling of phone duration; intra-word transformations using special rules for transforming phones depending on phonetic context and lexical stress [6]; rate-dependent phone sets [6]; speaking rate dependent (or adapted) acoustic models [1][8][9]; temporal wrapping in front-end processing, such as the continuous frame rate normalization (CFRN) technique [10]; variable frame rate (VFR) analysis [11]; etc.

A reliable estimation of speaking rate is often needed in order to adapt the ASR system to variations in speaking rate. Existing rate of speech (ROS) estimators can be characterized by two properties: (1) the metric (or measure) and (2) the duration of the speech segment on which ROS is estimated (for example, per speaker, per utterance, per word, etc.). The most common speech rate metrics are: word rate [6] (number of words per minute), phone (or syllable) rate. Normalization and phone duration percentile [6] are used for more accurate and robust estimations. All these measures are

A. Ronzhin et al. (Eds.): SPECOM 2014, LNAI 8773, pp. 418–424, 2014.

based on linguistic information. They can be calculated from segmentation (if reference transcriptions are available) or from the recognizer output. These approaches give reliable estimations of speaking rate in case of high quality segmentation. When the accuracy of the ASR system is low, then other methods are to be used. An alternative approach is to directly estimate speaking rate from the speech signal. An example of a signal-based measure is the mrate measure [2], which incorporates multiple energy-based estimators. In [3] Gaussian Mixture Models (GMM) were trained to estimate speech rate categories. In the same work, an ANN model was used as a mapping function from output likelihoods of these GMMs to continuous estimation of the speaking rate.

In this work we investigate a novel approach to speaking rate estimation based on Deep Neural Networks (DNNs). DNNs have recently been used with great success in automatic speech recognition tasks [13], and it is interesting to explore their application to the speaking rate estimation problem.

2 Speaking Rate Estimation

As mentioned before, one of the most common ways to measure ROS is calculating the number of phones per second. There are several variations of this formula [1], such as the Inverse of Mean Duration (IMD), where the total number of phones in the unit is divided by the total duration of the unit in seconds; or the Mean of Rates (MR) formula. These measures do not take into account the fact that mean durations of different phones and triphones in the database may vary a lot. To get more accurate estimations, we use normalization on the average length of triphones, computed from the training database. Thus the measure of speaking rate (or ROS) is defined as follows:

$$\rho = \frac{n}{\sum_{i=1}^{n} l(i)/\overline{l}(i)}, \tag{1}$$

where n is the number of phones in the unit; $l(i)$ is the actual duration of the triphone i in the unit; $\overline{l}(i)$ is the expected (mean) duration of triphone i, estimated from the training database.

An important question is at which level this formula should be calculated. These levels include per speaker, per utterance or per word calculation. The choice depends on the task and the type of speech. In [10] it was found that per utterance adaptation to speaking rate is more effective than per speaker adaptation for broadcast transcription tasks. In conversational speech, fluctuations in speech rate may be very high within one utterance, and per word ROS estimation may be more appropriate [6].

An example of how speech rate may vary from word to word in spontaneous conversation is shown in Figure 1. We can see than even within one short utterance speaking rate fluctuations are high.

A more detailed analysis of changes in the speaking rate is presented in Figure 2, where normalized histograms for relative differences in speaking rate between two adjacent words are shown. These histograms are calculated on a database of conversational telephone Russian speech. The relative difference is calculated for every pair of adjacent words in the database that satisfies following conditions: (1) the pause between

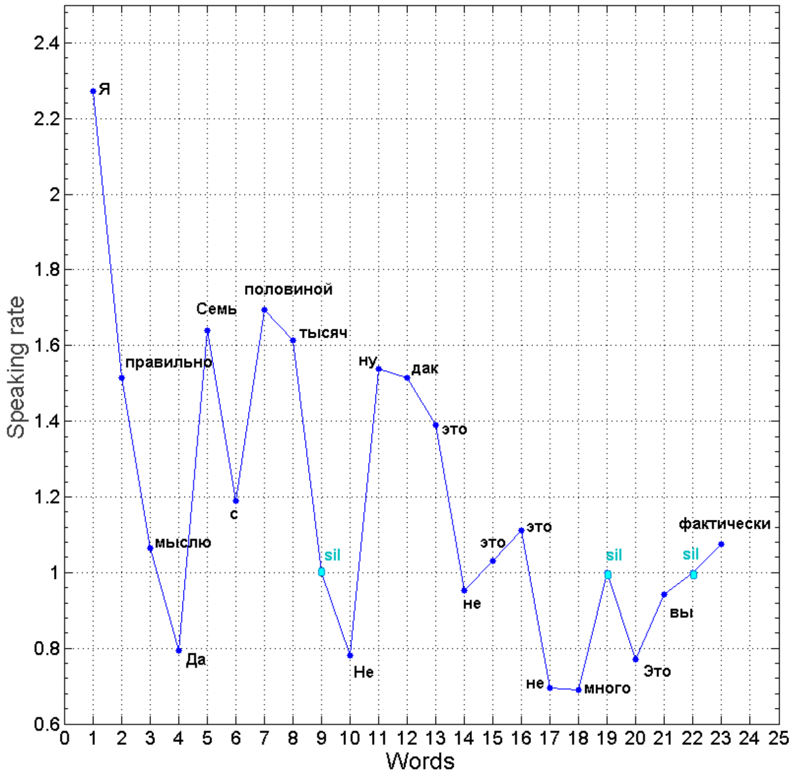

Fig. 1. An example of speaking rate estimation calculated for a spontaneous speech fragment

the two words does not exceed 250 ms; (2) the number of phones in these words is not smaller than a chosen threshold (1, 3 or 5 phones in our experiments). The relative difference $\Delta\rho$ in the speaking rate for two words with ROS ρ_1 and ρ_2 is calculated as follows:

$$\Delta\rho = \frac{2|\rho_1 - \rho_2|}{|\rho_1 + \rho_2|} \tag{2}$$

Mean and standard deviation statistics for relative difference $\Delta\rho$ in the speaking rate for two adjacent words are given in Table 1. We can see from Figure 1 and Table 1 that speaking rate is more stable for long words than for short ones. Note that ROS estimation for short words is less reliable than for long ones.

Based on the ROS estimation described above, we formed three groups, corresponding to slow, medium and fast speaking rate:

$$\text{ROS group} = \begin{cases} slow, & \text{if } \rho < 0.91 \\ medium, & \text{if } 0.91 \leq \rho \leq 1.11 \\ fast, & \text{if } \rho > 1.11 \end{cases} \tag{3}$$

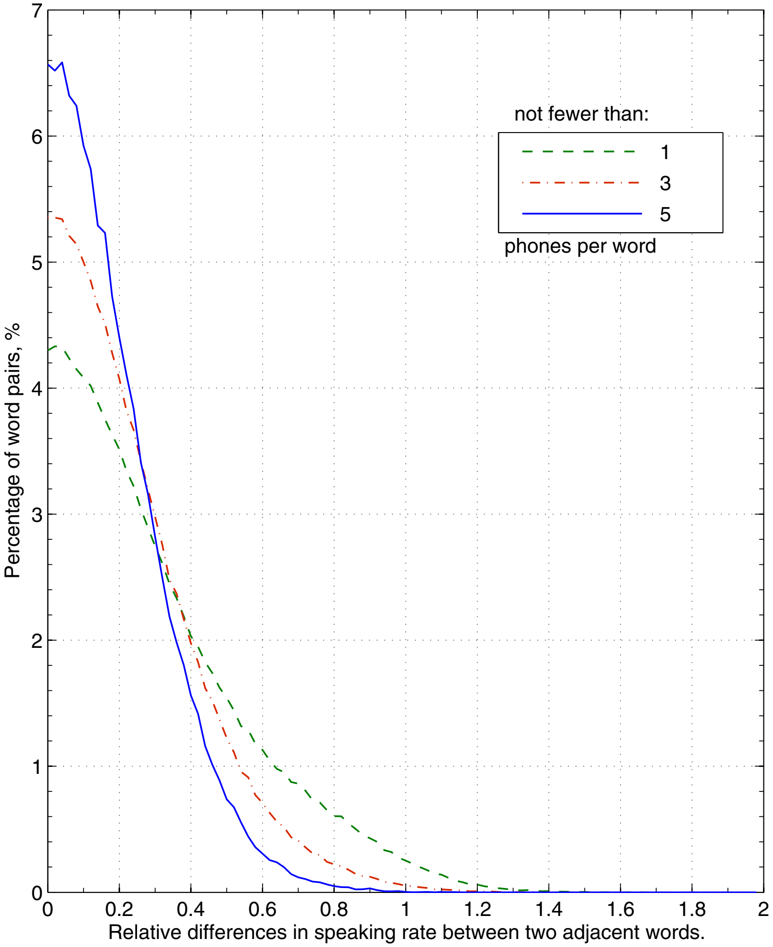

Fig. 2. Normalized histograms for relative differences in speaking rate between two adjacent words

3 DNN for ROS Estimation (ROS-DNN)

For the sake of simplicity, we train the DNN to classify each frame of speech into one of four classes: (1) slow, (2) medium, (3) fast speech or (4) silence. This classification may be extended further, since speaking rate is a continuous measure.

The training database was segmented into the four classes as follows. First, phone segmentation based on text transcripts was performed for each sentence in the training

Table 1. Mean and standard deviation statistics for relative difference $\Delta\rho$ in the speaking rate for two adjacent words

Minimum number of phones per word	Mean	Standard deviation
1	0.320	0.257
3	0.247	0.196
5	0.196	0.153

database. Second, ROS was estimated for each word in the segmented database, and the word was marked as slow, medium or fast according to its ROS value (as in Equation (3)); all pauses were marked as silence.

Thus, we had four DNN targets. The input features for the DNN were 13Δ and $13\Delta\Delta$ Mel-frequency cepstrum coefficients (MFCCs) spliced into 31 frames (± 15), resulting in 806-dimensional feature vectors. These features indicated the speaking rate dynamic.

The DNN system used three 1000-neuron hidden layers and one 4-neuron output layer. The DNN system was trained with layer-by-layer pre-training using the frame-level cross entropy criterion. The output layer was a soft-max layer.

4 Adaptation to Speaking Rate

Many HMM ASR systems have a fixed step size, typically 10 ms, in frontend processing. But it was shown in [10,11,14] that this approach is not optimal, as the best choice of frame rate may be affected by speaking rate. For example, for steady speech segments we can use a longer step size, while for rapidly changing segments the step should be shorter. This phenomenon was used in several approaches for adaptation to speaking rate, such as CFRN and VFR.

We performed adaptation to speaking rate simply by changing the step used for feature processing, based on the ROS estimation.

5 Experimental Results

We evaluated the proposed algorithm on a database of spontaneous conversational telephone Russian speech.

The DNN acoustic model for the ASR system was trained on approximately 270 hours of speech data. A subset from this training database was used to train ROS-DNN. Input features for training DNN acoustic models are 13-dimensional MFCC, spliced into 31 frames (± 15), resulting in 403-dimensional feature vectors. The same context length was used when processing input features for training ROS-DNN models. The DNN acoustic model had six hidden layers each with 1000 nodes and logistic activation, and an output layer with softmax activation. The output layer of the DNN acoustic model had a dimension of approximately 10K.

Table 2. Speech rate adaptation resuls for different ROS estimators: (1) the proposed approach based on using DNN models; (2) ground truth estimation based on ideal segmentation; (3) segmentation using texts derived from the recognizer

	Baseline	Rate-Adapted		
		DNN	Segmentation	
			Original transcripts	Texts from recognizer output
WER, %	38.3	37.4	37.0	37.6
Correcr, %	66.1	67.2	67.1	67.1

Experiments were conducted on approximately 300 recordings of conversational telephone speech. The total number of words in the test set was 3K.

We performed adaptation to speaking rate by changing step depending on ROS. We used a simple scheme for changing step: fast - 7.5 ms; medium - 9.4 ms; and slow - 10 ms.

A 200K vocabulary and a 3-gram language model were used in evaluation.

The performance results are shown in Table 2. The three columns on the right correspond to Word Error Rate (WER) results after adaptation based on three ROS estimators: (1) using DNN; (2) using phone segmentation as the ground truth estimation, and (3) estimation derived from the force alignment of recognition results.

We can see that adaptation based on all estimators leads to improved recognition performance: (1) 2.3% relative WER reduction for ROS-DNN, (2) 3.4% relative WER reduction for ROS based on phone segmentation, and (3) 1.8% relative WER reduction for ROS based on recognition results. Note that the acoustic models were trained with a fixed step (10 ms), and retraining them with rate-dependent step may improve results.

6 Conclusions

In this paper we propose a novel method for ROS estimation for the purpose of speaking rate adaptation. The proposed approach is based on training a DNN model on targets representing slow, medium and fast speech. Experiments with speaking rate adaptation show that our method of ROS estimation gives an improvement over unadapted baseline models. Moreover, the results of adaptation using the ROS-DNN estimator are better than the results of adaptation using the ROS estimator based on the recognition results. The advantage of this approach is that it does not require additional decoding passes.

This work presents only preliminary results of using DNNs for the task of speaking rate estimation. We believe that these results may be further improved by two modifications: first, by re-training acoustic models with a rate-dependent step; second, by using more groups for rate categorization or treating speaking rate as a continuous measure.

Acknowledgements. This work was partially financially supported by the Government of the Russian Federation, Grant 074-U01.

References

1. Mirghafori, N., Fosler, E., Morgan, N.: Towards robustness to fast speech in ASR. In: Proc. of the IEEE International Conference in Acoustics, Speech, and Signal Processing, ICASSP 1996, pp. 335–338 (1996)
2. Morgan, N., Fosler-Lussier, E.: Combining multiple estimators of speaking rate. In: Proc. of the IEEE International Conference In Acoustics, Speech, and Signal Processing, ICASSP-1996, pp. 729–732 (1998)
3. Faltlhauser, R., Pfau, T., Ruske, G.: On-line speaking rate estimation using gaussian mixture models. In: Proc. of the IEEE International Conference In Acoustics, Speech, and Signal Processing, ICASSP 2000, pp. 1355–1358 (2000)
4. Pfau, T., Ruske, G.: Estimating the speaking rate by vowel detection. In: Proc. of the IEEE International Conference In Acoustics, Speech and Signal Processing, ICASSP 1998, pp. 945–948 (1998)
5. Mirghafori, N., Foster, E., Morgan, N.: Fast speakers in large vocabulary continuous speech recognition: analysis & antidotes. In: Proc. of the EUROSPEECH, pp. 491–494 (1995)
6. Siegler, M.A.: Measuring and compensating for the effects of speech rate in large vocabulary continuous speech recognition (PhD Thesis). Carnegie Mellon University, Pittsburgh (1995)
7. Wrede, B., Fink, G.A., Sagerer, G.: An investigation of modelling aspects for rate-dependent speech recognition. In: Proc. of the INTERSPEECH, pp. 2527–2530 (2001)
8. Ban, S.M., Kim, H.S.: Speaking rate dependent multiple acoustic models using continuous frame rate normalization. In: Proc. of the Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC), Asia-Pacific, pp. 1–4 (2012)
9. Nanjo, H., Kato, K., Kawahara, T.: Speaking rate dependent acoustic modeling for spontaneous lecture speech recognition. In: Proc. of the INTERSPEECH, pp. 2531–2534 (2001)
10. Chu, S.M., Povey, D.: Speaking rate adaptation using continuous frame rate normalization. In: Proc. of the IEEE International Conference in Acoustics Speech and Signal Processing (ICASSP), pp. 4306–4309 (2010)
11. Zhu, Q., Alwan, A.: On the use of variable frame rate analysis in speech recognition. In: Proc. of the 2000 IEEE International Conference in Acoustics Speech and Signal Processing (ICASSP 2000), pp. 1783–1786 (2000)
12. Benzeghiba, M., De Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., ... Wellekens, C. Automatic speech recognition and speech variability: A review. Speech Communication 49(10), 763–786 (2007)
13. Hinton, G., Deng, L., Yu, D., Dahl, G.E., Mohamed, A.R., Jaitly, N., ... Kingsbury, B.: Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. Signal Processing Magazine 29(6), 82–97 (2012)
14. You, H., Zhu, Q., Alwan, A.: Entropy-based variable frame rate analysis of speech signals and its application to ASR. In: Proc. of the IEEE International Conference on In Acoustics, Speech, and Signal Processing – ICASSP 2004, vol. 1, pp. 549–552 (May 2004)