Integrating Prosodic Information into Recurrent Neural Network Language Model For Speech Recognition

Tong Fu, Yang Han, Xiangang Li, Yi Liu, Xihong Wu
Speech and Hearing Research Center,
Key Laboratory of Machine Perception (Ministry of Education),
Peking University, Beijing, 100871
{fut, hany, lixg, liuy, wxh}@cis.pku.edu.cn

***Abstract*—Prosody is a kind of cues that are critical to human speech perception and comprehension, so it is plausible to integrate prosodic information into machine speech recognition. However, as a result of the supra-segmental nature, it is hard to integrate prosodic information with conventional acoustic features. Recently, RNNLMs have shown to be the state-of-the-art language model in many tasks. We thus attempt to integrate prosodic information into RNNLMs for improving speech recognition performance based on rescoring strategy. Firstly, three word-level prosodic features are extracted from speech and then passed to RNNLMs separately. Therefore RNNLMs predict the next word based on prosodic features and word history. Experiments conducted on LibriSpeech Corpus show that the word error rate decreases from 8.07% to 7.96%. Secondly, prosodic information is combined on feature-level and model-level for further improvements and word error rate decreases 4.71% relatively.**

I. Introduction

Recently deep neural network has become one of the most popular methods in almost every field. The hybrid context dependent (CD)[1] deep neural network (DNN) hidden Markov model (HMM) (CD-DNN-HMM) becomes the dominant framework for speech recognition[2]. As a result of the ability of modeling complicated correlations in speech features, the CD-DNN-HMM performs much better than the conventional Gaussian mixture model (GMM) HMM.

However, there is still a gap between human speech recognition and machine speech recognition. Actually, human can integrate different information hidden in speech, such as speakers age, gender, mention, attitude, intention and so on from his/her voice, regardless of what is said. Prosody is such important information we need in speech perception processing[3]. For machine speech recognition system, a direct idea to integrate prosodic features is to combine them with other acoustic features when constructing acoustic models. However, prosody is actually related to various levels of information, from linguistic, para-linguistic, to non-linguistic and prosodic features spread to a wider range out of a phone or a syllable[4]. Its acoustic presentation is thus rather complicated which makes it hard to incorporate prosodic information into speech recognition systems.

Although the difficulties mentioned above, there are lots of studies trying to employ prosodic information to improve machine speech recognition performance. Some promising work includes the use of prosodic features to improve duration modeling[5], for controlling the search space and cross-word context models[6], to improve noise robustness recognition[7], to help to drive dynamic pronunciation modeling[8], and language models[9].

As mentioned above, it is difficult to integrate prosody information in acoustic models. However, we notice that the prosody information can implicitly reflect the emotion of speakers and can help human to predict what the speaker would say. Consider about this, in this paper, the prosody information is extracted and integrated with language models.

In recent years, neural network based language models (NNLMs)(feed-forward[10][11] or recurrent[12]) have shown success in both perplexity and word error rate (WER) compared to the conventional N-gram language models. The main reason is that the discrete nature of N-gram language models makes generalization a challenge while the NNLMs embed words in a continuous space. Therefore, the NNLMs can achieve better generalization for unseen N-grams. Moreover, NNLMs are easy to extend by adding extra input information such as prosodic features. When it comes to comparing feed-forward with recurrent, the recurrent connections allow the recurrent NNLMs to use arbitrarily long history while the feed-forward NNLMs are limited to fixed context. In this paper, we integrate word-level prosodic features into RNNLMs to improve speech recognition performance based on lattice rescoring method. The remainder of the paper is organized as follows. Section 2 reviews the structure of the typical RNNLM. In Section 3, we depict the main prosodic features we used and integrates prosodic information into RNNLMs. Next, in Section 4, we describe experiments for speech recognition and results. We discuss and conclude in Section 5.

II. The RNNLM

By using recurrent connections, information can cycle inside networks for arbitrary long time. In other words, RNNLMs have the ability of encoding temporal information implicitly for contexts with arbitrary length. The typical structure of a RNNLM can be described by a simple recurrent neural network or Elman network[13]. The structure of this recurrent neural network is described in Fig. 1.

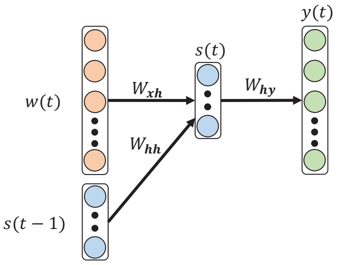

Fig. 1. Simple recurrent neural network.

It has an input layer x, a hidden layer s (also called the context layer or state of the network) and an output layer y. Input to the network at time t is $x(t)$, output is $y(t)$ and $s(t)$ is the state of the network (hidden layer). Input $x(t)$ is formed by concatenating $w(t)$ and $s(t-1)$, which represent the current word and output from the context layer(hidden layer) at time $t-1$ respectively. The computation among input, hidden and output layers are as follows:

$$x(t) = [w(t), s(t-1)] \tag{1}$$

$$s_j(t) = f\left(\sum_i x_i(t)u_{ji}\right) \tag{2}$$

$$y_k(t) = g\left(\sum_j s_j(t)v_{kj}\right) \tag{3}$$

where $f(z)$ is sigmoid activation function and $g(z)$ is softmax function.

By denoting the weight matrix of input $w(t)$ to hidden $s(t)$ as W_{xh}, input $s(t-1)$ to hidden $s(t)$ as W_{hh}, and hidden $s(t)$ to output $y(t)$ as W_{hy}. We can rewrite the computation equations as:

$$s(t) = f\left(W_{xh}w(t) + W_{hh}s(t-1)\right) \tag{4}$$

$$y(t) = g(W_{hy}s(t)) \tag{5}$$

Input vector $x(t)$ represents word at time t encoded using 1-of-N coding and previous context (hidden) layer, the size is equal to size of vocabulary V plus the size of the context (hidden) layer s. Output layer $y(t)$ represents probability distribution of next word given previous word $w(t)$ and context $s(t-1)$. Softmax function in output layer ensures that this probability distribution is valid, i.e., $y_m(t) > 0$ for any word m and $\sum_k y_k(t) = 1$.

III. Prosodic features and Integration

A. Feature extraction

The most frequently used acoustic correlates of prosody include fundamental frequency(or pitch), energy, and duration(or timing)[8]. We focus on features related to these three main cues. Pitch is an important prosodic feature and it might be changed when speakers emphasize words or express some emotions. As a result that speech is not strictly periodic, pitch will be varied with the time of opening and closing glottis. Pitch tracking is thus a tough task and influenced by many elements. One of most common approach for detecting pitch is NCCF (Normalised Cross Correlation Function, NCCF)[14][15]. We use the pitch extraction algorithm provided in KALDI Toolkit for frame level features[16], which is detailed in literature[17]. The resulting feature is a 3-dimensional vector consisting of pitch, NCCF and POV (Probability of Voicing, POV)[17]. In order to obtain the word level features, we average the frame level features within the word based on alignments.

Energy is a representation of the amplitude and the energy parameters are obtained by computing log RMS (Root Mean Square, RMS) energy for each frame speech feature, then frame level energy features are also averaged for the word level energy feature via alignments. For the j-th frame of speech, containing N sampling points $a_{j1}, \ldots, a_{jN}$, the log RMS energy is estimated by:

$$RMS_j = \log \sqrt{\frac{1}{N}\sum_{i=1}^{N} e_{ji}^2} \tag{6}$$

where e_{ji}^2 is the squared amplitude of a_{ji}.

Duration information is obtained by summing based on the alignments in training procedure, while the first time decoding results in testing procedure.

B. Integrating prosodic information into RNNLM

We can extend the basic model structure with extra input information to improve the model. As the extension method mentioned in literature[18], we mean to integrate prosodic features into RNNLM and structure is shown in Fig. 2. The extra input vector $p(t)$ represents prosodic features of word $w(t)$ at time t. So the input vector $x(t) = [w(t), s(t-1), p(t)]$, and the output vector $y(t)$ represents the probability distribution over word history from the vocabulary given the word $w(t)$, the context vector $s(t-1)$ and the prosodic feature vector $p(t)$.

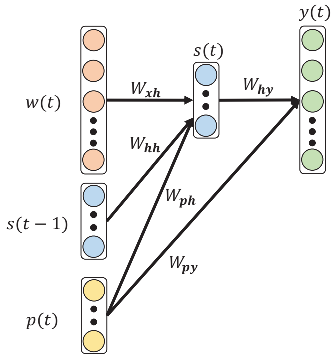

Fig. 2. Recurrent neural network with prosodic features $p(t)$.

The corresponding equations are modified as follows:

$$s(t) = f(W_{xh}w(t) + W_{hh}s(t-1) + W_{ph}p(t)) \tag{7}$$

$$y(t) = g(W_{ph}s(t) + W_{py}p(t)) \tag{8}$$

The training of this RNNLM consists of optimizing the weight matrices W_{xh}, W_{hy}, W_{hh}, W_{ph}, W_{py}. The algorithm of training refers to literatures[10][19].

IV. Experiments

In this paper, we train and validate all the models on a new corpus of reading English speech – LibriSpeech corpus [12]. The LibriSpeech corpus contains about 1000 hours of speech sampled at 16 kHz and derived from audiobooks that are part of the LibriVox project which is currently responsible for the creation of approximately 8000 public domain audio books and the majority are English. The training portion of the corpus is split into three subsets, with approximate size 100, 360, 500 hours respectively. Meanwhile, the speakers in the corpus are ranked according to the WER (Word Error Rate, WER) by a recognition system which is composed of a bigram language model and an acoustic model trained by corpus publisher on WSJs si-84 data subset[12], i.e. the lower-WER speakers are designated as "clean" and the higher-WER speakers are designated as "other". The details of the corpus partitions are shown in the literature[20].

A. Experimental setups

We build a GMM-HMM based speech recognizer for attaining the N-best list where the GMM-HMM is trained on about 960 hours training data from the corpus marked with "train-clean-100", "train-clean-360", "train-clean-500" and the test set is about 5 hours from "test-clean" in literature[20]. The speech is represented with 25ms frames of MFCCs (Mel-frequency Cepstral Coefficients, MFCCs), along with their first- and second-order temporal derivatives.

A 3-gram language model is estimated using all the acoustic model training transcriptions and the size of the N-best list on test set is 1000. The WER and perplexity of the 3-gram are showed in TABLE I. There are 84894 words in the vocabulary we use in experiments, so the representation of the current word input is 84894-dimensional. We use one hidden layer with 300 hidden units to remember the word history, therefore the context input is 300-dimensional. In the experiments, we use the "dev-clean" data as the development set and the learning rate for training each RNNLM is decreased exponentially, and the initial and final learning rates are set specific to each network for stable convergence of training. The number of layers we unfold when conducting BPTT (Back-propagation Through Time, BPTT) training is set to 4 for all RNNLMs, which refers to literature[19]. When it comes to the baseline model, we use a conventional RNNLM, without any prosodic feature inputs, to rescore the N-best list. In the experiment for baseline RNNLM, the weight of RNNLM is 0.7 and 3-gram language model is 0.3 which are derived from best results of development set. The baseline has 85194 (84894+300) input units and 300 units in hidden layer. The initial learning rate is 0.1. The WER of the baseline is 8.07%.

B. Integrating single different prosodic features into RNNLM

Prosodic features are compared by building RNNLMs integrated with three different prosodic feature separately. The WERs and perplexities are summarized in TABLE I.

TABLE I
WERs and PPLs of 3-gram, RNNLM and prosodic RNNLMs.

Model Descriptions	WER(%)	PPL
Baseline: 3-gram	10.81	308.09
Baseline: RNNLM	8.07	231.78
RNNLM+Pitch	7.95	211.59
RNNLM+Energy	7.97	212.25
RNNLM+Duration	8.08	214.43
RNNLM+PED	8.03	213.95
RNNLM-INTER	7.69	-

The "Baseline: RNNLM" represents the conventional RNNLM, "RNNLM+Pitch" means the RNNLM integrated with pitch features. "RNNLM+Energy" represents the RNNLM with energy input and the energy feature is only 1-dimensional. "RNNLM+Duration" is the RNNLM integrated with duration information and its dimensionality is also one. In order to compare the baseline RNNLM with other proposed RNNLMs, the weight of 3-gram language model is fixed on 0.3 for all the experiments. TABLE I shows that "RNNLM+Duration" achieves a WER of 8.08% which is almost equal to the WER achieved by baseline, so the introduction of duration information independently seems to have no effect on improvements. However, the integration with pitch features works well and the "RNNLM+Pitch" model gets a WER of 7.95% which means 1.49% relative improvements over baseline. Meanwhile, the energy information also works and "RNNLM+Energy" model achieves a WER of 7.97%, almost the same improvements with "RNNLM+Pitch". In case of perplexities, "RNNLM+Pitch", "RNNLM+Energy" and "RNNLM+Duration" models all reduce the perplexity.

C. Integrating combined features into RNNLM

As a result that different prosodic features might be relevant and they may help each other to provide more efficient information, we explore to combine prosodic information for further improvements. There are two candidate ways to combine different prosodic information, i.e., feature-level combination and model-level combination. The feature-level one is to combine different features together to form a "bigger" feature, while the model-level combination is to conduct interpolation of individual RNNLMs mentioned above, which are integrated with single different prosodic features. We validate these two ways and the results are shown in TABLE I.

The model marked with "RNNLM+PED" is constructed by combining all the three prosodic features into a bigger feature which is then used as the extra input for the RNNLM. Meanwhile, "RNNLM-INTER" means we utilize the prosodic information through conducting interpolation of the three models mentioned in the previous part, i.e., “RNNLM+Pitch”, “RNNLM+Energy”, “RNNLM+Duration”, and the weights of “RNNLM+Pitch”, “RNNLM+Energy” and “RNNLM+Duration” are 0.3, 0.3 and 0.1 separately, also derived from the development set. We find that the feature-level combination does not work well as we expected although it achieves improvements both in WER and PPL. While the WER of “RNNLM-INTER” is 7.69%, which means a 4.71% relative improvement.

V. Discussion and Conclusions

From the experimental results given in “RNNLM-INTER”, we can observe the improvement is larger than the sum of individual improvements, which implies that the gains due to each individual prosodic RNNLM are somewhat complementary, but not fully additive as expected. It also suggests that different prosodic information are in fact correlated. Individual prosodic RNNLMs trained with single kind of prosodic features seem can preserve the effectiveness of individual prosodic feature and different RNNLMs catch prosodic information on different levels at first. Then the interpolation method, which can be viewed as a vote mechanism, makes RNNLMs cooperate with each other to “vote” for decisions that improve the gains further. Moreover, it is interesting to find that there is no relation between WER and PPL of “RNNLM+Duration” model compared with baseline. One of the possible reasons is that the duration information is more sensitive to the alignments which are hard to be precise sometimes.

Language models are aimed to provide a predictive probability distribution for the next word conditioned on the words seen so far. In addition to the previous words, prosodic information in the audio stream, which is one kind of parallel knowledge source to the word(or text) stream, can be used as complementary information for predicting the next words in language models. From a multimodal learning perspective[21], the use of audio information(prosody) refines the text condition(word), we think this is one of the reasons why prosodic information improved the RNNLMs as the experimental results showed, and is also the initial motivation for this paper.

Prosodic patterns of spontaneous speech are more various while those of reading speech are consistent relatively. Therefore, prosodic patterns in reading speech are likely to be more accessible for modeling. Although experiments are conducted on reading speech and gain improvements, we believe that prosody is more important and helpful for spontaneous speech tasks, and more experiments will be conducted on spontaneous speech data for the future work.

In this paper, different word-level prosodic features are integrated into RNNLMs and speech recognition performance improves based on the rescoring method. For further improvements, we explore to employ prosodic information combinations. Methods of feature-level combination and model-level combination are validated. The model-level combination method achieves a 4.71% relative improvement.

Acknowledgment

The work is supported in part by National Basic Research Program (973 Program) of China (No. 2013CB329304), the National Natural Science Foundation of China (No. 90920302, No. 91120001, No.61121002), the “Twelfth Five-Year” National Science & Technology Support Program of China (No. 2012BAI12B01) and the Key Program of National Social Science Foundation of China (No. 12&ZD119).

References

[1] M. Hwang and X. Huang, “Shared-distribution hidden markov models for speech recognition,” *Speech and Audio Processing, IEEE Transactions on*, vol. 1, no. 4, pp. 414–420, 1993.

[2] G. Hinton, L. Deng, D. Yu, G.E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T.N. Sainath, et al., “Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups,” *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 82–97, 2012.

[3] M. Ostendorf, I. Shafran, and R. Bates, “Prosody models for conversational speech recognition,” in *Proc. of the 2nd Plenary Meeting and Symposium on Prosody and Speech Processing*, 2003, pp. 147–154.

[4] C. Wang, *Prosodic modeling for improved speech recognition and understanding*, Ph.D. thesis, Massachusetts Institute of Technology, 2001.

[5] V.R. Gadde, “Modeling word duration for better speech recognition,” in *Proceedings of NIST Speech Transcription Workshop*, 2000.

[6] S. Lee, K. Hirose, and N. Minematsu, “Incorporation of prosodic modules for large vocabulary continuous speech recognition,” in *ISCA Tutorial and Research Workshop (ITRW) on Prosody in Speech Recognition and Understanding*, 2001.

[7] K. Iwano, T. Seki, and S. Furui, “Noise robust speech recognition using f0 contour extracted by hough transform,” in *ICSLP-2002*, 2002, pp. 941–944.

[8] J.E. Fosler-Lussier, *Dynamic pronunciation models for automatic speech recognition*, Ph.D. thesis, University of California, Berkeley Fall 1999., 1999.

[9] K. Hirose, N. Minematsu, and M. Terao, “Statistical language modeling with prosodic boundaries and its use for continuous speech recognition,” in *INTERSPEECH*, 2002, pp. 937–940.

[10] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin, “A neural probabilistic language model,” *The Journal of Machine Learning Research*, vol. 3, pp. 1137–1155, 2003.

[11] H. Schwenk, “Continuous space language models,” *Computer Speech & Language*, vol. 21, no. 3, pp. 492–518, 2007.

[12] T. Mikolov, M. Karafiát, L. Burget, J. Cernockỳ, and S. Khudanpur, “Recurrent neural network based language model.,” in *INTERSPEECH*, 2010, pp. 1045–1048.

[13] J.L. Elman, “Finding structure in time,” *Cognitive science*, vol. 14, no. 2, pp. 179–211, 1990.

[14] D. Talkin, “A robust algorithm for pitch tracking (rapt),” *Speech coding and synthesis*, vol. 495, pp. 518, 1995.

[15] Md N Ibne Nazrul, MT Hossain Setu, Shahed Hussain, and Md Kumrul Hasan, “An effective speech preprocessing technique for normalized cross-correlation pitch extractor,” in *ISSPIT*. IEEE, 2003, pp. 749–752.

[16] D. Povey, A. Ghoshal, L. Burget, and others., “The kaldi speech recognition toolkit,” *in ASRU*, pp. 1–4, 2011.

[17] P. Ghahremani, B. BabaAli, D. Povey, K. Riedhammer, J. Trmal, and S. Khudanpur, “A pitch extraction algorithm tuned for automatic speech recognition,” in *ICASSP*. IEEE, 2014, pp. 2494–2498.

[18] T. Mikolov and G. Zweig, “Context dependent recurrent neural network language model.,” in *SLT*, 2012, pp. 234–239.

[19] T. Mikolov, “Statistical language models based on neural networks,” *Presentation at Google, Mountain View, 2nd April*, 2012.

[20] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: an ASR corpus based on public domain audio books,” in *ICASSP*. IEEE, 2015.

[21] Songfang Huang and Steve Renals, “Using prosodic features in language models for meetings,” in *Machine Learning for Multimodal Interaction*, pp. 192–203. Springer, 2008.