Prosodically-enhanced Recurrent Neural Network Language Models

Siva Reddy Gangireddy[1], Steve Renals[1], Yoshihiko Nankaku[2] and Akinobu Lee[2]
s.gangireddy@sms.ed.ac.uk

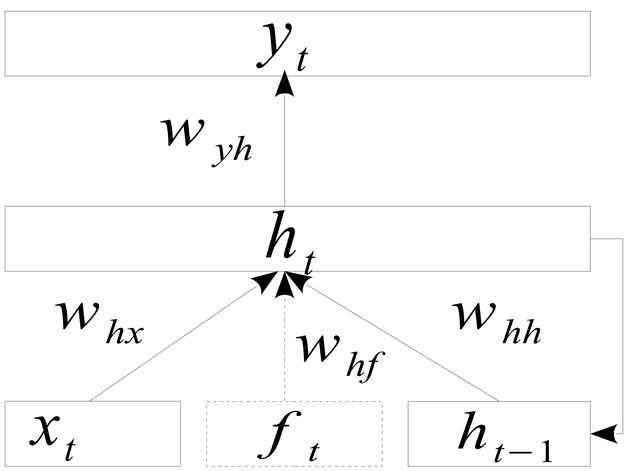
[1]Centre for Speech Technology and Research, University of Edinburgh, UK
[2]Nagoya Institute of Technology, Nagoya, Japan

Motivation

- Current ASR systems model the acoustics at sub-word (phone) level
- Prosody features are suprasegmental and can't easily fit into this framework
- Prosody features has rich additional information and can help to improve the speech recognition accuracies
- Prosody information is available at various levels
  - With in the words (word and phone duration)
  - Between the words (pause duration)
  - Beyond the words (F0 contour)
- In this work we have trained the RNNLMs to jointly model the words and their prosodic features
- It is more natural to capture the prosody at syllable level (Huang and Renals, 2007)
- We have also trained the RNNLMs on prosodic features computed at syllable level

RNNLM with Feature Layer

- More natural to model the continuous prosody with neural language models than with *n*-grams
- We have added an extra feature layer to jointly model the words and related prosodic features



Recurrent neural network language model with a feature layer

Prosody Features

- **Word** duration and **pause** duration between the words are obtained by force aligning the transcripts with their acoustics
- **Final phone duration**
  - Pause duration effect the duration of the preceding word (pre-pausal lengthening)
- **Syllable duration and F0**
  - Automatic syllable extraction algorithm used to extract the syllables (Aylett 2006)
  - Simple alignment procedure used to obtain the syllables preceding the current word
  - Four F0 features are computed at each syllable: mean, maximum, minimum and range of F0

Text Experiments

- RNNLMs are trained on Switchboard training transcripts
- Training & Validation : 3.2M & 130K tokens
- Testing: 20 Switchboard conversations of *eval2000*
- *n*-grams: Switchboard and Fisher transcripts

Model	Validation	+KN3	*eval2000-swbd*	+KN3
KN3	82.4	82.4	81.9	81.9
RNNLM	78.4	70.6	77.5	70.8
RNNLM-pause	68.6	64.4	66.5	63.9
RNNLM-worddur	70.8	65.8	76.7	68.2
RNNLM-fphonedur	70.2	65.3	70.7	66.1
RNNLM-syldur	63.7	61.4	65.0	62.0
RNNLM-sylF0	70.1	64.6	67.3	63.6

Perplexities of 3-gram, RNNLM and prosody RNNLMs

ASR experiments

- **Switchboard task**
  - Evaluated on 20 Switchboard conversations of *eval2000*
  - GMM and DNN-based acoustic models are used to force align and compute the prosody features
- **TED task**
  - RNNLMs are trained on combination of TED lectures and AMI data
  - Evaluated on *tst2011*
  - DNN acoustic models are used to force align and compute the prosody features
  - Acoustic models are trained on TED and AMI data

Switchboard Task

- Significant improvements with **word duration** model and moderate improvements with other models

Model	GMM(%WER)	DNN(%WER)
3-gram	19.5	19.5
RNNLM	18.1	18.1
RNNLM-pause	18.0	17.9
RNNLM-worddur	17.6	17.9
RNNLM-fphonedur	17.8	18.0

%WERs computed on 100-best lists of *eval2000* data set (Switchboard conversations only)

TED Task

- The models are not effective enough to improve the WERs

Model	PPL	%WER
3-gram	120.2	12.6
RNNLM	198.0	11.9
RNNLM-pause	184.1	12.0
RNNLM-worddur	194.1	11.8
RNNLM-fphonedur	184.2	11.7

PPLs and %WERs computed on *tst2011* and 100-best lists of *tst2011*, respectively

Syllable Duration and F0

- Models are trained by varying the number of syllables in the context of the current word
- 0.3% absolute improvement with **RNNLM-syldur5** model, trained on five syllables in the context
- Not observed any improvements with models trained on syllable F0 features

Model	PPL	%WER
3-gram	81.9	19.5
RNNLM	77.5	18.1
RNNLM-syldur3	63.5	17.9
RNNLM-syldur5	65.0	17.8
RNNLM-syldur10	63.2	18.0

%WERs are computed on 100-best lists of *eval2000* data set (Switchboard conversations only).