Chinese-English Mixlingual Automatic Speech Recognition System

Emily Hua, Kelly Chen, Wendy Wang
Department of Computer Science, Engineering, QMSS @ Columbia University in the City of New York

Introduction

Our project aims to build a Mandarin-English mixlingual ASR.
The mixlingual speaking habit brings a code-switching phenomenon, where more than one language occurs within an utterance. This is very common in many multilingual countries.
For example, the following sentence involves an intra-sentence code-switching:

Deep learning project presentation 今晚 due

which means "Deep learning project presentation is due tonight". We want to tackle this code-switching challenge using hybrid models.
Traditional ASR uses GMM-HMM system, which assumes Gaussian distribution of speech signals as associated with each of HMM states. However, this fixed distribution assumption is not necessarily true. Neural Nets come in to learn features and provides posterior for decoding without assuming any particular structure of the data. Our project shows hybrid systems improves the traditional GMM-HMM system by 10% absolute on this Mandarin-English Code-Switching in South-East Asia (SEAME) mixlingual dataset we acquired from Linguistic Data Consortium (LDC).

GMM-HMM

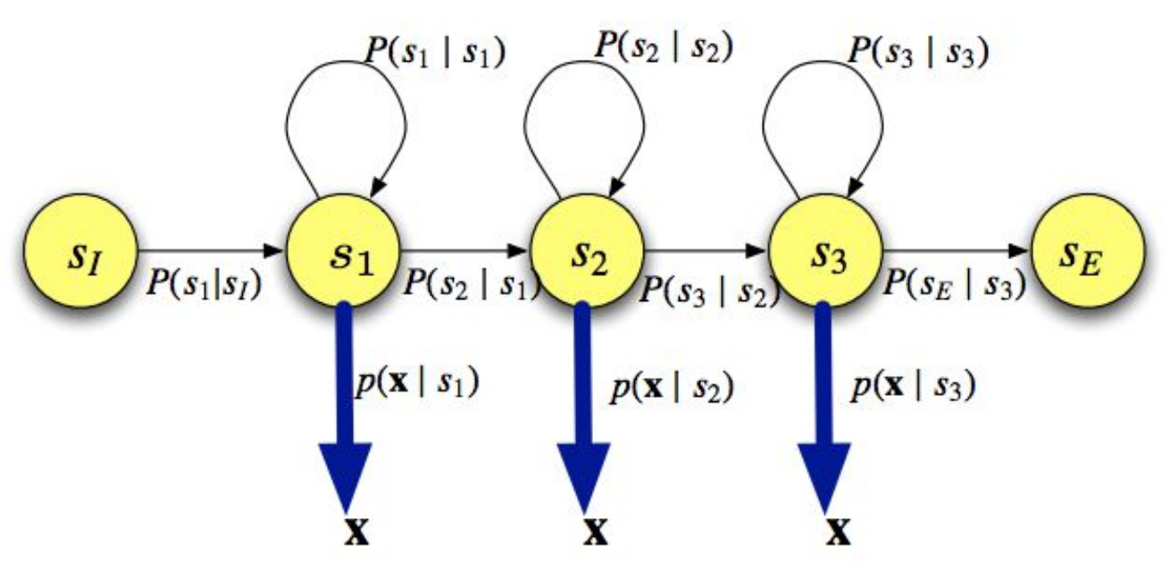


Fig 1. HMM Parameters:
$a_{kj} = P(s_j|s_k)$,
where a_{kj} is the transition probability from state k to state j.
$b_j(x) = p(x|s_j)$,
where b_j is the emission probability from state j to sequence X.
GMM provides posterior b_j.

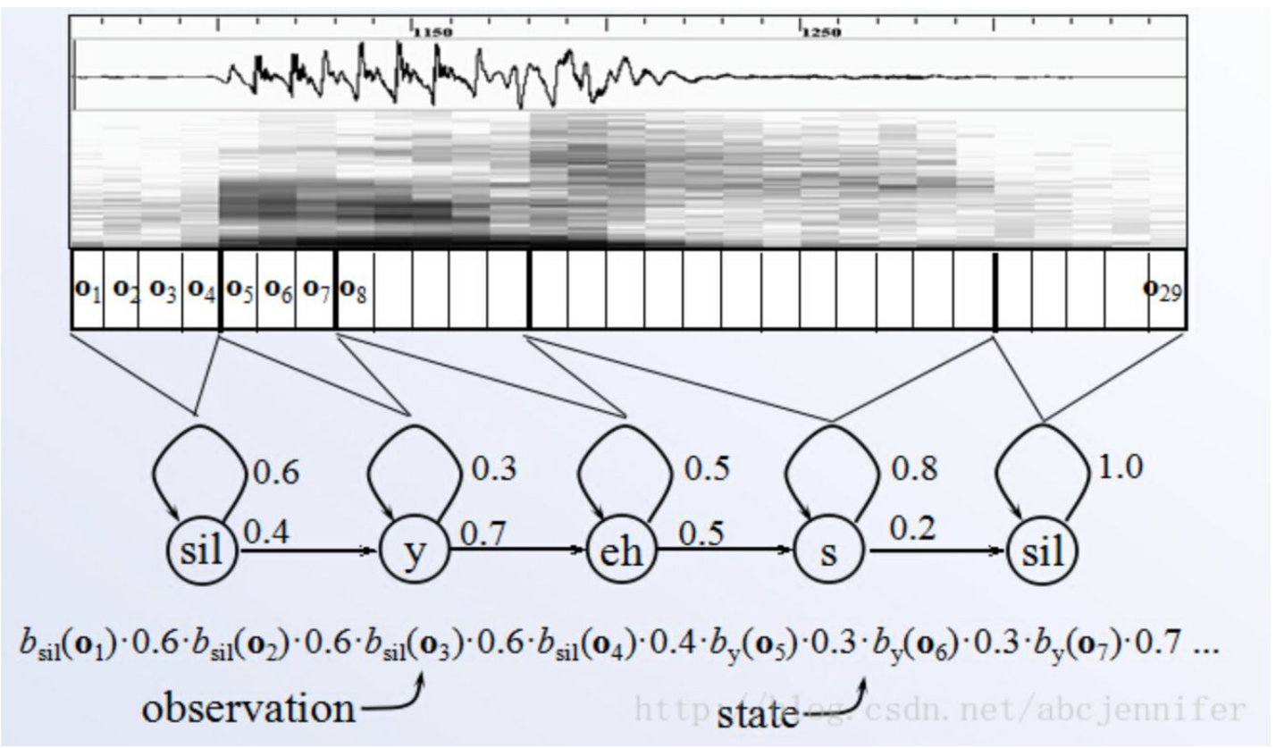
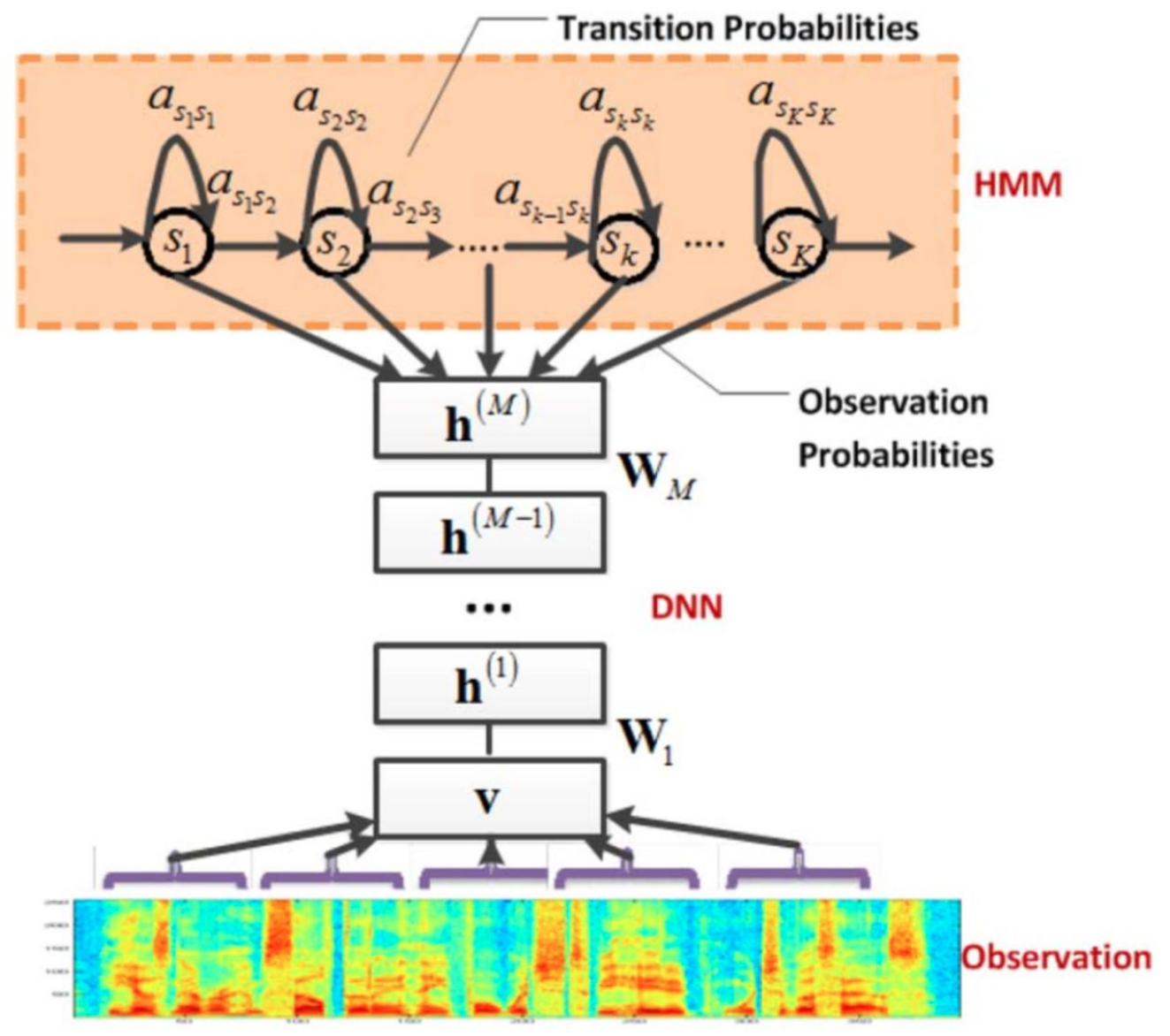

o_1 o_2 o_3 o_4 o_5 o_6 o_7 o_8 o_{29}
sil
y
eh
s
sil
0.6
0.4
0.3
0.7
0.5
0.5
0.8
0.2
1.0
$b_{sil}(\mathbf{o}_1)\cdot 0.6\cdot b_{sil}(\mathbf{o}_2)\cdot 0.6\cdot b_{sil}(\mathbf{o}_3)\cdot 0.6\cdot b_{sil}(\mathbf{o}_4)\cdot 0.4\cdot b_y(\mathbf{o}_5)\cdot 0.3\cdot b_y(\mathbf{o}_6)\cdot 0.3\cdot b_y(\mathbf{o}_7)\cdot 0.7$ ...
observation
state
http://blog.csdn.net/abcjennifer

Fig 2. GMM-HMM illustration:
1) cut waveform into frames with equivalent length,
and then extract MFCC features from each frame.
2) run GMM on each frame to get $b_j(o_i)$ associated with each frame o_i
3) use transition probability a_{kj} and emission probability b_j to calculate the probability that a frame being generated by a state sequence,
whichever sequence has the highest probability,
the model will output it as the best path.

Hybrid System



In the hybrid model, instead of using GMM, we use Deep Neural Nets to estimate posterior probabilities. The output of the neural network is the probability of a phone class given the feature $P(s_j|x)$. In order to compute the emission probability $P(x|s_j)$ ($b_j(x)$) for HMM, we uses the Bayes Rule: $P(x|s_j) = P(s_j|x) \times P(x)/P(s_j)$, which can be simplified as $P(x|s_j)\ P(s_j|x)/P(s_j)$ (this is okay, as $P(x)$ does not depend on the class s_j). This being said, we scaled the neural net output by class priors $P(s_j)$, in order to get the emission probability for HMM.

Fig 3. Hybrid System Diagram.
Here GMM-HMM creates the forced alignment between features and phone states. These phone states served as targets for the Deep Neural Networks, while the inputs to DNN are the usual features. The Neural Network provides the conditional probability, which can be scaled as emission probability for the HMM.

Data Preprocessing (Transcript Clean up) & Data Description

Acoustic Data Information:

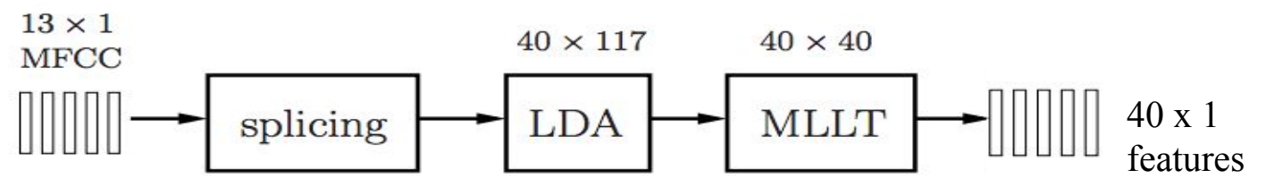
46 hours of audio. 159 unique speakers in total, 14 speakers in test set. 46613 utterances in total, 3392 utterance in test set.

Language Data Information:

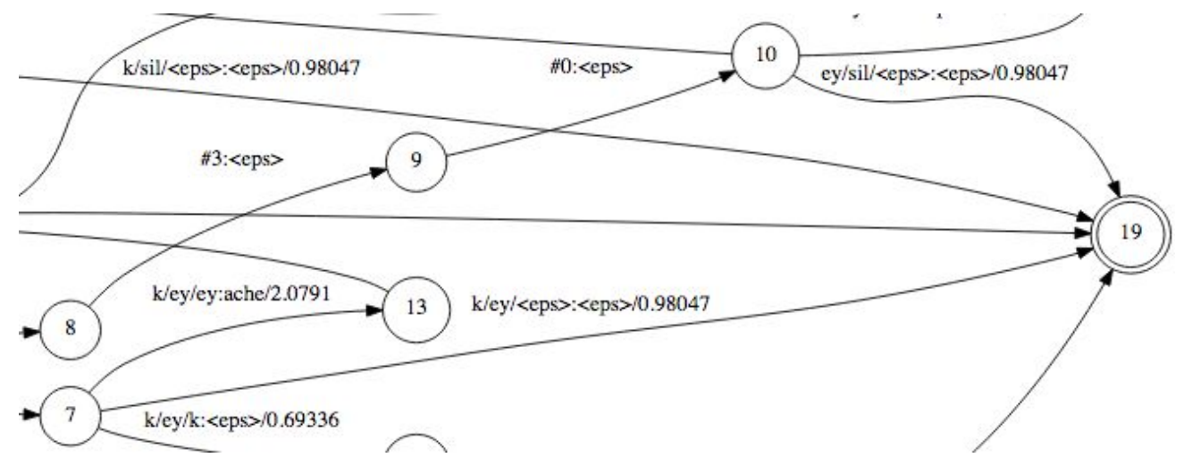
Training set: English oov 1244/9560 = 0.13
Training set: Chinese oov 831/3683 = 0.23

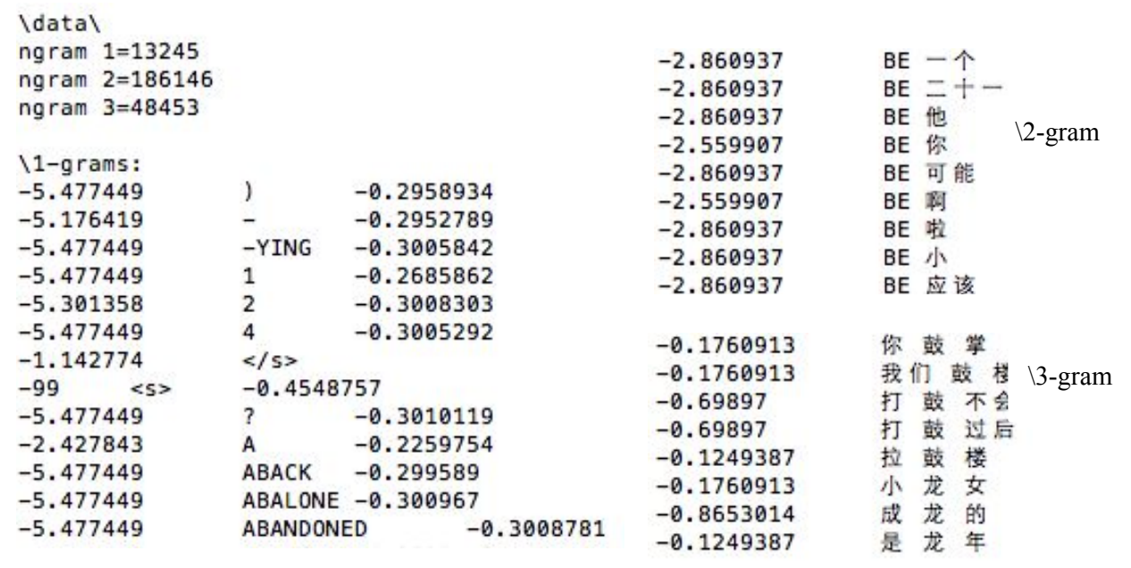
The raw transcripts were preprocessed according to following rules:
- transform fullwidth forms to halfwidth forms
- split Chinese characters and English words if concatenated without space (e.g., sorry我不吃 -> sorry 我不吃)
- correct misspelled phrases (e.g., abit -> a bit)
- remove all annotation signs that are not explained in documentation, including % and " (e.g., %chelsia% -> chelsia)
- remove annotation sign for word of foreign language in case the word exists in cmu dictionary (e.g., #sushi# -> sushi)
- fix wrong annotation (e.g., [ppl] -> (ppl), ppl -> (ppl))
- replace annotations with SIL index (e.g., [oh] -> SIL2)
- remove single period (.)
- remove other punctuation, like ? and)
- remove utterances having unsegmented Chinese characters (length of Chinese segment > 4, like 哦那个我还不能啦)
- split Chinese OOV words into characters to reduce OOV rate

Acoustic Features & Language Model



Our baseline features are obtained as follows:
- Mel-frequency cepstral coefficients(MFCC) and its cepstral mean and vairance normalization is applied on a per speaker basis.
- The resulting 13-dimensional features are spliced in context size of 7 frames (i.e. ±3),
- followed by de-correlation and dimensionality reduction to 40 using linear discriminant analysis (LDA).
- The resulting features are further decelerated using maximum likelihood linear transform (MLLT), which is also known as global semi-tied covariance (STC).
We followed Kaldi recipe on Voxforge to generate the aforementioned features.





```
\data\
ngram 1=13245
ngram 2=186146
ngram 3=48453

\1-grams:
-5.477449       )        -0.2958934
-5.176419       -        -0.2952789
-5.477449       -YING    -0.3005842
-5.477449       1        -0.2685862
-5.301358       2        -0.3008303
-5.477449       4        -0.3005292
-1.142774       </s>
-99      <s>    -0.4548757
-5.477449       ?        -0.3010119
-2.427843       A        -0.2259754
-5.477449       ABACK    -0.299589
-5.477449       ABALONE -0.300967
-5.477449       ABANDONED        -0.3008781

-2.860937       BE 一个
-2.860937       BE 二十一
-2.860937       BE 他
-2.559907       BE 你           \2-gram
-2.860937       BE 可能
-2.559907       BE 啊
-2.860937       BE 啦
-2.860937       BE 小
-2.860937       BE 应该

-0.1760913      你 鼓 掌
-0.1760913      我们 鼓 楼      \3-gram
-0.69897        打 鼓 不会
-0.69897        打 鼓 过后
-0.1249387      拉 鼓 楼
-0.1760913      小 龙 女
-0.8653014      成 龙 的
-0.1249387      是 龙 年
```

Fig 4. Snippet of unigram, bi-gram, tri-gram in our language model.
We used a tri-gram language model based on the unique words showing up in our training dataset.

Fig 6. HCLG.fst
a fully expanded decoding graph (HCLG) that represents the language-model, pronunciation dictionary (lexicon), context-dependency, and HMM structure in our model. The output is a Finite State Transducer that has word-ids on the output, and pdf-ids on the input (these are indexes that resolve to Gaussian Mixture Models).

DNN

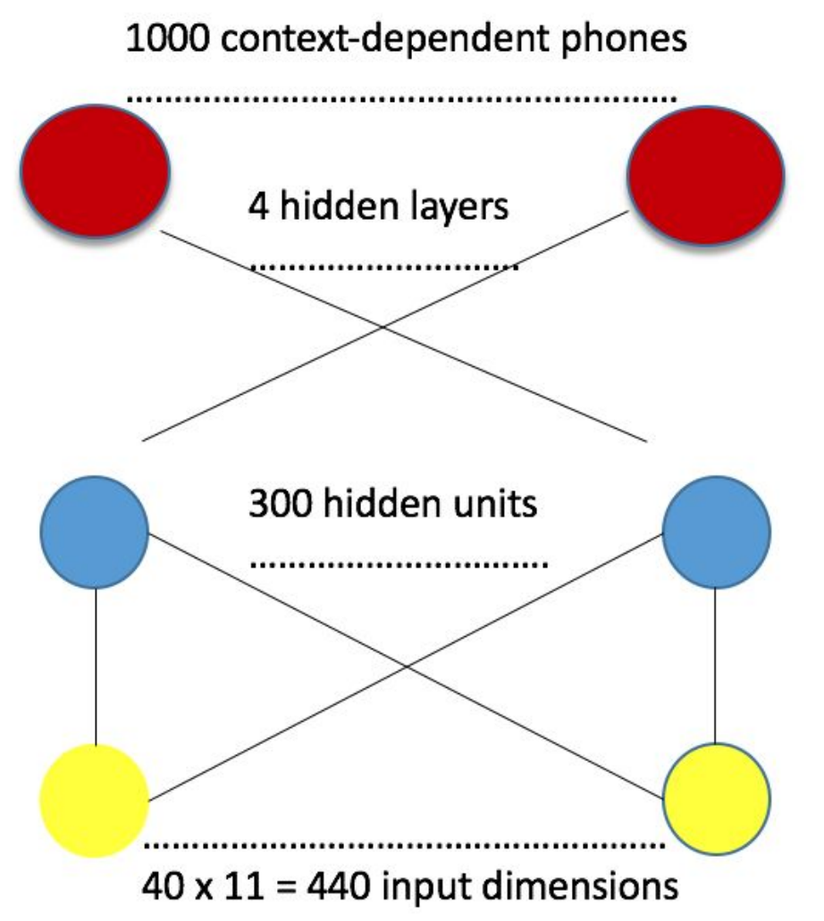


Fig 5. Baseline Network Structure
In our network p is set to 2 (a.k.a Euclidean Norm), and we have 4 hidden layers, mini-batch size of 512, 8 epochs, 40 (LDA) x11 input dims (audio feature), 1000 output dims (context-dependent triphones), 300 for hidden layer output dims.

Network Description:
We used a Maxout (*p-norm*) Network for our baseline DNN.
p-norm is a "dimension-reducing" non-linearities.
p-norm:

$$y = ||x||_p = (\sum_i |x_i|^p)^{1/p} \quad ,$$

where the vector x represents a small group of input.
The performance of p-norm outperforms rectified linear units(ReLU) and tanh units [1].
There is also a "normalization layer" that scales down "the whole set of activations if necessary to prevent the standard deviation from exceeding 1."[2]
σ is the uncentered standard deviation of x_i:

$$\sigma = \sqrt{1/K \sum_i (x_i)^2} \quad ,$$

where x_i is the input.
The nonlinearity is:

$$y_i = \begin{cases} x_i, & \sigma \leq 1 \\ x_i/\sigma, & \sigma > 1 \end{cases}$$

This is applied directly after *p-norm* (without a layer of weights in-between) to stabilize unbounded-output nonlinearities.

Results

Table 1. Effect of Transcript Cleanup on WER under LDA-MLLT Acoustic Model

Modification	WER
Chinese OOV split into characters	65.90%
Chinese OOV split into characters + Merge SIL	**64.27%**
Chinese OOV split into characters + English OOV fixed/lexicon inserted	67.59%
Chinese OOV split into characters + Merge SIL + English OOV fixed/lexicon inserted	66.33%

- LDA-MLLT model is trained on alignment of first triphone pass, which is trained and aligned on the result of monophone system.
- To our surprise, fixing English OOV words and inserting lexicons reduces WER.

Table 2. Comparison of GMM-HMM Models and Hybrid Models

Feature	GMM-HMM WER	Hybrid WER	Hybrid Details
MFCC	64.27%	54.89%	DNN
MFCC + pitch	63.54%	52.82%	DNN
FBank	63.84%	54.19%	CNN
FBank + pitch	63.69%	running	CNN

- DNN model is from Kaldi nnet2 recipe.
- CNN model is from Kaldi nnet recipe.
- Our best model is hybrid DNN using MFCC and pitch features, with 52.82% WER and 45.66% CER.

Further Thoughts

We are going to check the following aspects to improve the performance, if time permits:

Transcript Improvement
- change British English to American English as in the CMU lexicon (e.g., specialise − > specialize)
- fix misspelled English words (e.g., avalable − > available)
Lexicon Dictionary Improvement
- combine simple words in lexicon to reduce OOV rate (e.g., hand + phone − > handphone)
- modify pronunciation in lexicon according to linguistic rules (as mentioned in [3])
Audio Data Augmentation
- change the speed of the audio signal, after which an average relative improvement of 4.3% in WER was reported. [4]
- vocal tract length perturbation, or VTLP, after which an average improvement of 0.65% in PER in DNN and that of 1.0% in CNN was reported. [5]

Acoustic Model ReAlignment
- currently we are using a relatively naive acoustic model (monophone − > first triphone pass − > LDA_MLLT)
- an example shows that keep training and realigning can further reduce the WER by 15% on the basis of that (... − > FMLLR − > SAT − > SGMM)
Language Model Improvement
- smoothing technique on n-gram language model, like Kneser-Ney or Good-Turing
RNN LM: RNN do not make the Markov assumption and so can, in theory, take into account long-term dependencies when modeling natural language. The main advantages would be the greater representational power of neural networks and their ability to perform intelligent smoothing by taking into account syntactic and semantic features.

References

Fig 1.
University Of Edinburgh-ASR Course Slides, 2013
Fig 2.
GMM-HMM语音识别模型-原理篇, Rachel Zhang, 2014
Fig 3.
Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition, Dahl et al., 2011
Fig 6.
Decoding graph construction in Kaldi: A visual walkthrough, Vassil Panayotov, 2012
Some Kaldi Notes, Josh Meyer, 2016
[1][2]
Improving Deep Neural Network Acoustic Models Using Generalized Maxout Networks, Povey et al, 2014
[3]
A First Speech Recognition System For Mandarin-English Code-switch Conversational Speech, Vu et al., 2012
[4]
Audio Augmentation for Speech Recognition, Ko et al., 2015
[5]
Vocal Tract Length Perturbation (VTLP) improves speech recognition, Jaitly & Hinton, 2013