Modeling Context and Language Variation for Non-Native Speech Recognition

Tien-Ping Tan, Laurent Besacier

LIG/GETALP Laboratory, UMR CNRS 5217
UJF - BP 53, 38041 Grenoble Cedex 9, France
tien-ping.tan@imag.fr, laurent.besacier@imag.fr

Abstract
Non-native speakers often face difficulty in pronouncing like the native speakers. This paper proposes to model pronunciation variation in non-native speaker’s speech using only acoustics models, without the need for the corpus. Variation in term of context and language will be modeled. The combination of both modeling resulted in the reduction of absolute WER as much as 16% and 6% for native Vietnamese and Chinese speakers of French.

Index Terms: non-native ASR, context modeling, language modeling, interpolation, merging

1. Introduction
Automatic speech recognition applications are becoming increasing popular. However, as automatic speech recognition matured, speech recognition performance on non-native speakers is still low.

Non-native speech has the characteristics of slower speaking rate, broader distribution compare to native speech, pronunciation mistakes, smaller working vocabulary and disfluency. Non-native speakers face difficulty in articulating the target language phonemes. For new phonemes which are not found in the speaker’s mother tongue, it is a challenge for beginners at least to articulate unfamiliar phonemes. On the other hand for similar phonemes which exist in both the target language and the speaker’s native language, they may have trouble changing certain articulation habits which are specific to their mother tongue [1][2].

Current automatic speech recognition systems take advantage of context clues for accurate recognition. As a result, non-native speakers are often unable to profit from it. On the contrary, it may even end up deteriorating their recognition. So, for non-native speakers, we may want to revisit the context dependent acoustic modeling concept to reduce the sharpness of the context and combine the features of both the speaker’s native language (L1) and non-native speech (L2) to model these variations.

Getting non-native speech for acoustic modeling is often difficult and in some cases unfeasible. Therefore, research in acoustic model adaptation attempts to use limited non-native speech or the speaker’s mother tongue to improve the target language acoustic models. The proposed method needs only the target language acoustic models and the speaker’s native language acoustic model to model context and language features in non-native speakers.

This paper is organized as following. In Section 2, we present related works in non-native’s acoustic modeling. In Section 3, we describe our approach in modeling context and language variations. Section 4 gives the experimental results and finally, conclusions are drawn in Section 5.

2. Related Works
In general, three common approaches for non-native acoustic modeling are acoustic model reconstruction (re-training), acoustic model interpolation and acoustic model merging.

2.1. Acoustic Model Reconstruction
Through acoustic model training, a new model tailored to non-native speakers can be created. Current ASR systems which use sharp triphone context dependent (CD) information do not perform very well with non-natives. On the other hand, a context independent (CI) model or a context dependent model with smaller and more appropriate shared state [3] may turn out to suit non-native speakers better. Non-native speaker’s language cues can also be fused into target language acoustic model by training it with some non-native speech or the speaker’s native language [4]. Another interesting approach is to incorporate language variations through state tying and interpolation during acoustic model reconstruction [5][6].

2.2. Acoustic Model Interpolation
Instead of modeling variation between two models using acoustic model reconstruction, acoustic model interpolation can also be used to produce the effect. Acoustic model interpolation can be performed by estimating a-priori weights to be multiplied to two or more acoustic models involved. The nearest corresponding Gaussian from source state (constructed from speaker’s native language (L1) [7] or constructed from a small amount of non-native speech (L2) [8]) is selected for interpolation for every Gaussian in target state using certain distance measure. Other variants can also be found [9].

2.3. Acoustic Model Merging
In acoustic model merging, non-native speaker’s language features are represented by combining one or more corresponding models from target and source (can be L1 or L2) acoustic to form a new model [7][10]. The transition weights can be estimated automatically or manually. For example in Figure 1 below, a target model /s/ and a source model /s_1/ are combined to create a new model /p/ with six states. Transition weights $w_1$ and $w_2$ can be assigned manually or automatically to the target and source models.

Figure 1. A merged model.
3. Modeling Context and Language Features

Acoustic model merging and interpolation are interesting approaches to model pronunciation variation which exists among non-native speakers in a simple and fast manner. We propose in this paper a hybrid approach of merging and interpolation to model context and language variations.

The general approach of interpolation is to select the nearest corresponding Gaussian from source state for every Gaussian in target state using certain distance measure. Instead, we propose to carry out interpolation in a different manner, where every Gaussian in target state is treated like the ‘centroid’ for the Gaussians in source state. The subsequent step is to find the nearest associated centroid or target Gaussian for all the source Gaussians using distance measure like Euclidean distance or approximated divergence distance.

Euclidean distance = \( (\Sigma (\mu_i - \mu_j)^2)^{1/2} \)
Approximated divergence distance = \( (\Sigma (\mu_i - \mu_j)^2 / (\sigma_i \sigma_j))^{1/2} \)

Every source Gaussian will be associated with only one target Gaussian. Certain target Gaussians will be instead associated with zero or more source Gaussians. When the distance between the associated target Gaussian and the source Gaussian is below a threshold, their means, variances and mixture weights will be interpolated (case 1). Otherwise, merging is performed: for those target Gaussians without any associated source Gaussian (case 2) or for the source Gaussian that are far (more than the threshold) from their associated target Gaussians (case 2). In case 2 and 3, their mixture weights will be reduced by the interpolation weight. The threshold can be calculated for example by measuring the average distance among the Gaussians, and then multiplying it with a constant. The resulted model is a hybrid model of interpolation and merging.

\[
p_{\text{new,sn}} = w \cdot p_{\text{tg,sn}} + (1-w) \cdot p_{\text{sc,sn}}, \quad p_{\text{new,sn}} \neq \emptyset,
\]
\[
d(p_{\text{tg,sn}}, p_{\text{sc,sn}}) < \text{dist} \quad (1)
\]
\[
p_{\text{new,sn}} = p_{\text{tg,sn}}, \quad \omega_{\text{new,sn}} = (1-w) \cdot \omega_{\text{tg,sn}}, \quad p_{\text{new,sn}} \neq \emptyset,
\]
\[
d(p_{\text{tg,sn}}, p_{\text{sc,sn}}) > \text{dist} \quad (2)
\]
\[
p_{\text{new,sn}} = p_{\text{tg,sn}}, \quad \omega_{\text{new,sn}} = \omega \cdot \omega_{\text{tg,sn}}, \quad p_{\text{new,sn}} \neq \emptyset
\]
\[
(3)
\]

where \( p_{\text{new,sn}} \) = interpolated/merged Gaussian, \( p_{\text{tg,sn}} \) = target Gaussian, and \( p_{\text{sc,sn}} \) = source Gaussian. \( w \) = interpolation weight, \( 0 \leq w \leq 1.0 \), \( \omega \) is the mixture weight for the Gaussian. \( d() \) is a distance function and \( \text{dist} \) is a threshold distance

In cases where the models are context dependent model, the matching triphone context will be looked upon. If there is no matching triphone context, the context independent context will be used.

3.1. Language Variation

Research showed that there is an interaction between target language and native phonology system for non-native speakers [2]. The proposed method can be used to model language variation between a target language acoustic model (target model) and a speaker’s native language acoustic model (source model). Before the proposed method is carried out, the target state and their corresponding source states can first be matched using knowledge based approach such as IPA table or data driven methods like confusion matrix. Figure 2 below shows an example of what will take place.

![Figure 2. Acoustic space. \( p_{\text{FR, s1g1}} \) is French /p/ of state 1, Gaussian 1 and \( p_{\text{VN, s1g1}} \) is Vietnamese /p/ of state 1, Gaussian 1. \( p_{\text{FR, s1g2}} \) and \( p_{\text{VN, s1g2}} \) will be associated with \( p_{\text{PR, s1g1}} \). Both will be interpolated with \( p_{\text{PR, s1g1}} \), \( p_{\text{FR, s1g2}} \) and \( p_{\text{VN, s1g2}} \) which are far away from each other will be merged (both Gaussians are kept and their mixture weight values will be recalculated).](image)

3.2. Context Variation

Context independent model is a bit flat while context dependent model can be too sharp for non-native speakers. So, the approach discussed above can be used for modeling between contexts. The context can be modeled to reach an intermediate state between these two extremes. When modeling context variation, the model with a smaller number of states will be treated as the target model while the other will be considered as the source model. The process of modeling context variation is similar to modeling language variation discussed in previous section. One thing different is that since all models with bigger number of states also belong to the model with smaller number of states (both are from the same language), all source Gaussians are assumed to have a target Gaussian interpolation partner. So, no threshold needs to be set.

For example, if we have a CI model (target model) and a CD model (source model). All CD triphones are matched to their corresponding CI monophones. Next, the corresponding CI Gaussian for every CD Gaussian is found using certain distance measure. Interpolation is then performed on CI Gaussians with their associated CD Gaussians, while the CI Gaussian without any interpolation partner will be merged.

4. Experiments

The experiments were carried out on our non-native French corpus [11] using CMU Sphinx 3 ASR. There are two groups of non-native speakers: Chinese and Vietnamese. Each speaker read about a hundred sentences related to tourism domain. Baseline French Continuous HMM acoustic model was created from BREF120 corpus [12]. As for the source languages, we have a 15 hours Vietnamese corpus [13] and a 5 hours Mandarin Chinese corpus [14]. The general domain trigram language model was created using Le Monde newspapers text, and subsequently interpolated with a tourism domain language model (from NESPOLE project).

In most cases, the interpolated acoustic model will have different number of Gaussians per state. Currently Sphinx architecture is not capable of handling varied number of Gaussians per state. To model this, we set all states to the maximum number of Gaussians possible, the means, variances and mixture weights for the empty Gaussians are set to zeros.

4.1. Baseline Experiment Results

Some baseline tests which model the context and language features were prepared for comparison purpose against our proposed method.
4.1. Context Variation

Acoustic model with different number of tied-states were trained while maintaining the number of Gaussians at 16. Besides that, CI models with different number of Gaussians were also prepared. Below are the results:

Table 1. WER of non-natives using CD acoustic models at different number of tied-state, with 16 Gaussians

<table>
<thead>
<tr>
<th>State</th>
<th>CI: 43</th>
<th>429</th>
<th>629</th>
<th>4129</th>
<th>8129</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vietnamese</td>
<td>60.6</td>
<td>59.7</td>
<td>59.6</td>
<td>66.5</td>
<td>70.2</td>
</tr>
<tr>
<td>Chinese</td>
<td>58.5</td>
<td>63.8</td>
<td>67.5</td>
<td>78.0</td>
<td>83.0</td>
</tr>
</tbody>
</table>

Table 2. WER of non-natives using CI acoustic models with different number of Gaussians

<table>
<thead>
<tr>
<th>Gaussian</th>
<th>16</th>
<th>32</th>
<th>64</th>
<th>128</th>
<th>256</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vietnamese</td>
<td>60.6</td>
<td>57.4</td>
<td>56.4</td>
<td>55.8</td>
<td>54.9</td>
</tr>
<tr>
<td>Chinese</td>
<td>58.5</td>
<td>56.6</td>
<td>56.1</td>
<td>55.6</td>
<td>56.0</td>
</tr>
</tbody>
</table>

The average WER for native Vietnamese and Chinese speakers are high. The high difficulty of the database was confirmed by human perception tests\(^1\) which showed average WER of 12.1% and 11.3% respectively. The results show that sharp context modeling does not improve and even degrades compared to context independent (CI) modeling.

4.1.2. Language Variation

Before modeling language variation, we need to determine the matching phonemes in the target language and the speaker’s native language. We used a combination of knowledge based and data driven approaches to determine the target language phoneme substitution by non-natives speakers [9].

Acoustic model interpolation and merging were experimented. The interpolation described in section 2.2 was applied. For acoustic model merging, instead of using the architecture in Figure 1, we assumed it is possible to transit from a Gaussian in one state to another Gaussian in another corresponding state in another language and vice versa. So, each new model had only three states, where the corresponding states in target and source models were merged to become one. Transition weights to different models were instead applied to the mixture weights of the models.

Sixteen Gaussians CI French and the speaker’s native language (Vietnamese and Chinese) acoustic model were used in the experiment. The results showed that acoustic model merging performs better than interpolation in most cases.

4.2. Proposed Approach Experiment Results

This section presents the results of our proposed approach.

4.2.1. Context Variation

The context modeling was carried out using a 43 states CI model and an 8129 states CD model. Both have 16 Gaussians per state. The interpolation-merging produced a CD model with 8129 states with an average of 25 Gaussians per state (except for CI weight=0 and CI weight=1.0). Approximated divergence distance was used as the distance measurement.

We noticed that there was a slight decrease in WER when the CI weight is at 1.0, compare to the original CI result. Note that when CI weight equal to 1.0, the algorithm will produce a model with 8129 states, where all triphones replaced by their respective monophone. The best WER for native Vietnamese and Chinese speakers of French were achieved when CI weight was 0.7. The WER were 51.5 and 54.0 respectively.

We validated the result on another corpus, by conducting the test on 23 native German speakers of English from the corpus ISLE [15]. TIMIT [16] corpus was used to create CI and CD models with 1120 states. The models were then interpolate-merge to create a new CD acoustic model. The test showed that when CI weight is at 1.0, the WER for non-native speakers was 57.9%. On the other hand, the WER of CD acoustic model was 63.5%. At CI weight equals to 0.5, German speakers recorded a reduction in WER to 55.1%.

The results from our context modeling show that when appropriate weight is used, the hybrid method produces encouraging results with ASR. The weight to apply corresponds with the experience of the speaker in the language. The Vietnamese speakers which are more experienced show higher improvements in WER compare to Chinese speakers. We also found that the WER of native French speakers (not showed in the graph) only showed slight increase of 2% compared to CD model when the CI weight is equal to 0.5.

4.2.2. Language Variation

The approach can also be used to model language variation. A French CI acoustic model with 43 states Gaussian was

---

\(^1\) Human listeners were asked to transcribe non-native utterances, where an unlimited number of replays for each utterance were permitted.
interpolated-merged with the speaker’s native language CI acoustic model. Both have 16 Gaussians per state. Euclidean distance was used as the distance measure. The resulted models have an average of 26 Gaussians per state. Threshold was set at about two times the average Gaussian distance.

The test showed that the proposed hybrid method performed better than the interpolation approach. It is better than acoustic model merging in most cases and the results are achieved using less Gaussians compared to acoustic model merging. For example at FR weight equals to 0.5, the hybrid models have WER of 51.2% and 53.4 for Vietnamese and Chinese respectively, while in merging the WER are 53.6% and 55.7% respectively.

![Figure 5](http://www.d-ear.com/CCC/corpora.htm)

Figure 5. Graph shows interpolation-merging of French CI model and Vietnamese/ Chinese CI model.

### 4.2.3. Context and Language Variations

In this test, we investigate the combined effect of modeling the context and language variations. We used the CD model after our proposed context modeling at CI weight equals to 0.5. The CD model was then interpolated-merged with the speaker’s native acoustic model. The Vietnamese CD model which contains 5123 states and the Chinese CD model with 1102 states were used.

![Figure 6](http://www.d-ear.com/CCC/corpora.htm)

Figure 6. Combination of context and language modeling

At French (FR) weight of 0.5, the test showed an overall improvement of WER (compared to the baseline presented in Table 1) from 60.6% to 44.1% for Vietnamese speakers and from 58.5% to 52.1% for Chinese speakers of French. On the other hand, the WERs for native French speakers increased only about 3% with the modified Vietnamese and Chinese model, compared to the baseline CD model.

### 5. Conclusions

We have presented a hybrid approach of interpolation and merging to model context and language variations which can be used offline. By applying a weight of 0.5, the method will give generally good results. When non-native speech is available, weight can be estimated automatically. Since there is only one parameter, one possibility is to compare the acoustic scores of different acoustic models created at different weights and to select the one which produces the highest value. One side effect of the resulted hybrid model is that it increases the total number of Gaussians. Appropriate clustering of the Gaussians or eliminating certain unnecessary Gaussians may be useful to reduce the number of Gaussians and thus the overall complexity of the model.

### 6. References


