A Statistical Model of Korean Loanword Phonology

HAHN KOO

San José State University

The paper presents a statistical model of Korean loanword phonology which predicts the adapted form in Korean for a given word in English. In essence, the model is a first-order hidden Markov model, where the states represent Korean phoneme strings of variable length and emit English phonemes as observation. The parameters of the model are trained on pairs of English word and the corresponding form in Korean according to maximum likelihood estimate. The most likely sequence of Korean phonemes given an English phoneme string is found by applying the Viterbi algorithm. Performance of the model is evaluated on a subset of Korean loanword database compiled by the National Institute of the Korean Language. As the model incorporates virtually no linguistic knowledge, the evaluation results can be used as baseline performance for future research in developing linguistically sophisticated models.

1. Introduction

Recent studies in loanword phonology have addressed specific issues for which there is no simple phonological explanation such as whether the
neutral vowel /ɨ/ is inserted word-finally when an English word that ends with a stop is adopted in Korean (Kang 2004). Needless to say, such studies are crucial in developing a theory. However, another type of effort one should make in developing a theory is to assess its overall quality by implementing a model and testing it against reasonably large data. For example, if one were to develop a generative model of Korean loanword phonology in terms of rewrite rules of the form \( X \rightarrow Y / W_1 Z \), one should not only identify the phonological context in which /ɨ/ is inserted word-finally, but also assess how accurately the entire set of such rewrite rules explain a large collection of Korean loanwords.

To evaluate the overall quality of a model, one needs data and performance of other models to compare with. In response, the current study presents a simple statistical model whose performance on a database of Korean loanwords can be compared with those of more linguistically sophisticated models in the future. In brief, the model is a first-order hidden Markov model that predicts the most likely sequence of Korean phonemes given an English phoneme string. As we shall see, the model incorporates virtually no linguistic knowledge, so the performance level reported in this study can be used as baseline performance for comparison.

The paper is organized as follows. In Section 2, I present a statistical interpretation of the adaptation process in loanword phonology. In Section 3, I introduce a few simplifying assumptions I made to build a working model of the adaptation process. I describe how the idea is implemented as a first-order hidden Markov model in Section 4. The implemented model is evaluated on a Korean loanword dataset in Section 5. I conclude the paper in Section 6.

2. A statistical interpretation of the adaptation process

In this paper, I assume that predicting how the form of a word in the source language is adapted in the recipient language is equivalent to identifying the most likely adapted form in the recipient language given the source form. That is, predicting how a source form \( s \) is adapted in the recipient language where potential candidates are enumerated in a set \( R \) is equivalent to solving the equation in (1) and identifying what \( \hat{r} \) is.

\[
\hat{r} = \arg \max_{r \in R} P(r \mid s) \quad (1)
\]

By Bayes’ rule, solving (1) is equivalent to solving (2).

\[
\hat{r} = \arg \max_{r \in R} P(s \mid r) \cdot P(r) \quad (2)
\]
The two conditional probabilities $P(s|r)$ and $P(r)$ are often called the likelihood and the prior, respectively, in the computational linguistics literature. In this context, the likelihood $P(s|r)$ is the probability that the source form of a given candidate $r$ is $s$, as opposed to some other source forms. The prior $P(r)$ is the probability of observing $r$ as an adapted form that corresponds to some source form, whatever the source form may be.

For example, assuming that the source language is English and the recipient language Korean, the likelihood $P(/bɹaɪə/ | /pɪɹə/) \approx /pɹəɪə/ (‘briar’) is the probability that /pɹəɪə/ is how the English word /bɹaɪə/ (‘briar’) is adopted in Korean rather than some other English word. The prior $P(/pɪɹə\!/) \approx /pɪɹə\!$/ in a list of Korean loanwords. To predict how ‘large’ is adopted in Korean, one would calculate $P(/bɹaɪə/ | r) \cdot P(r)$ for every candidate form $r$ and choose the one that maximizes the product of the two probabilities.

By making several assumptions including the independence assumption and the bigram assumption, I solve the equation in (2) with a first-order hidden Markov model whose states represent component units of the recipient language and generate component units of the source language as their observation. The most likely candidate form can be identified by running the Viterbi algorithm on the given source form. Details of the assumptions and the hidden Markov model implementation are described in the next two Sections.

3. Assumptions

One major issue with the statistical approach in Section 2 is how to estimate the likelihood and the prior specified in (2). A straightforward approach would be to get their maximum likelihood estimates from pairs of source form and the corresponding adapted form in the recipient language. That is, the likelihood and the prior can be estimated according to (3) and (4) as follows, where $C(s,r)$ denotes the number of times $s$ is paired with $r$, $C(r)$ denotes the number of times $r$ appears as an adapted form, and $N$ denotes the total number of adapted forms.

$$P(s \mid r) = \frac{C(s, r)}{C(r)} \quad (3)$$

$$P(r) = \frac{C(r)}{N} \quad (4)$$
For example, suppose we had a database of Korean loanwords, like the one in Section 5, consisting of English words paired with their adapted form in Korean. The likelihood $P(\text{/baɾaiw/} \mid \text{/piraiʌ/})$ can be estimated by dividing how often /baɾaiw/ is paired with /piraiʌ/ by how often /piraiʌ/ appears in the database as the adapted form. The prior $P(\text{/piraiʌ/})$ can be estimated by dividing how often /piraiʌ/ appears in the database as the adapted form by the total number of adapted forms in the database.

However, the problem with (3) and (4) is that the frequencies $C(s, r)$ and $C(r)$ will be very small. It would be realistic to assume that each adapted form in the database appears at most once and is paired with just one word in the source language. That is, it is highly likely for most $r$ that $C(r)=1$, and $C(s, r)=1$ or even $C(s, r)=0$ for many pairs. As a result, using the maximum likelihood estimates of the probabilities as in (3) and (4) for prediction would be meaningless.

In response, for a given pair of $s$ and $r$, I rewrite each of the forms as a string of smaller component units and make the independence assumption and the bigram assumption to estimate the likelihood and the prior. Assuming that $s$ and $r$ are respectively rewritten as strings of $n$ component units $s_1s_2...s_n$ and $r_1r_2...r_n$, the likelihood and the prior can be approximated as in (5) and (6). Note that in approximating the prior, we first pad the adapted form with the word boundary symbols $<$s$>$ and $</s>$ at the beginning and the end, respectively. That is, $r_0=<s>$ and $r_{n+1}=</s>$ in (6).

$$P(s \mid r) \approx \prod_{i=1}^{n} P(s_i \mid r_i) = \frac{\prod_{i=1}^{n} C(s_i, r_i)}{C(r_i)}$$ (5)

$$P(r) \approx \prod_{i=0}^{n} P(r_{i+1} \mid r_i) = \frac{\prod_{i=0}^{n} C(r_i, r_{i+1})}{C(r_i)}$$ (6)

As long as the pair of component units $(s_i, r_i)$ and the bigram of component units of an adapted form $(r_i, r_{i+1})$ are frequently found in the given database, the approximated likelihood and the prior will be robust.

Thus, the final issue to be resolved is how to define and identify the component units of the source form and the adapted form. In this paper, I assume that the component units of the source form are its phonemes and that the component units of the adapted form are the sequences of zero or more of its phonemes that correspond to the component phonemes of the source form.

Note that the assumption I made for defining the component units of the adapted form requires alignment between the source form and the
adapted form. Given that both the source form and the adapted form are initially written as strings of phonemes, the forms can be aligned using dynamic programming after defining the costs of following three edit-operations: substituting a phoneme in the source language with a phoneme in the recipient language, inserting a phoneme in the recipient language after a phoneme in the source form, deleting a phoneme in the source form.

In the current study, the substitution cost is assumed to be inversely proportional to the phonological similarity between the two phonemes; the more similar the two phonemes are, the more expensive substituting one phoneme for the other is. To calculate the phonological similarity, every phoneme in both languages is specified in terms of features listed in Table 1 and the similarity between every phoneme pair is calculated according to Frisch et al. (1997) using the script by Albright (2003).

Table 1. List of phonological features in terms of which all phonemes in both the source language and the recipient language are specified.

<table>
<thead>
<tr>
<th>Phonological features</th>
</tr>
</thead>
<tbody>
<tr>
<td>consonantal, sonorant, continuant, strident, lateral, labial, coronal, dorsal, round, anterior, distributed, front, central, back, high, mid, low, diphthong, nasal, advanced-tongue-root, spread glottis, constricted glottis, voiced</td>
</tr>
</tbody>
</table>

The resulting similarity score lies between zero and one. The substitution cost is defined as one minus the similarity score. Insertion and deletion are assumed to be as bad as substitution that involves two phonemes with zero similarity. As a result, the insertion cost and the deletion cost are assumed to be one in this paper. Table 2 illustrates how /bɹɑɪəɹ/ in English and its adapted form /pɨɾɑiʌ/ in Korean are aligned using dynamic programming with the cost parameters thus defined.

Table 2. Illustration of how the phonemic transcription of the English word 'briar' is aligned with that of its adapted form in Korean.

<table>
<thead>
<tr>
<th>Language</th>
<th>Phonemic transcription</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>b ɹ a i r o j</td>
</tr>
<tr>
<td>Korean</td>
<td>p i r a i ã</td>
</tr>
</tbody>
</table>

The alignment example in Table 2 entails that the component units of the source form are its constituent phonemes /b/, /ɹ/, /ɑ/, /i/, /ɹ/, /o/, and /j/, while the component units of the adapted form are /pɨ/, /ɾ/, /a/, /i/, /ʌ/, and /NULL/. Note that the unit /pɨ/ consists of two phonemes and that NULL stands for a unit of zero length. In turn, the likelihood $P(\text{bɹɑɪəɹ} | /\text{pɨɾɑiʌ}/)$
is estimated by multiplying the following component likelihoods: $P(b|p\text{ɨ})$, $P(x|i)$, $P(\alpha|\alpha)$, $P(i|i)$, $P(\alpha|x)$, and $P(x|\text{NULL})$. The prior $P(p\text{ɨ}r\alphai\text{ʌ}|x)$ is the product over the following bigram probabilities: $P(p\text{ɨ}|<s>)$, $P(r'|p\text{ɨ})$, $P(\alpha|i)$, $P(i|\alpha)$, $P(\alpha|i)$, $P(\text{NULL}|x)$, and $P(</s>|\text{NULL})$. The maximum likelihood estimates of the component likelihoods and the bigram probabilities can be derived after all the pairs of source form and adapted form in the database are aligned.

4. Implementation as a first-order hidden Markov model

Once we have the likelihood and the prior, we need to identify the candidate form that maximizes the product of the two probabilities. One efficient way is to use a first-order hidden Markov model and the Viterbi algorithm.

The states of the model in the current study represent the set of constituent units in the recipient language, including the word-boundary symbols $<s>$ and $</s>$ described in Section 3. It should be clear that the set of constituent units can be identified by aligning all the pairs of source form and adapted form in a database of loanwords.

The model is in the initial state representing $<s>$ at the beginning. It probabilistically changes its state, possibly to the same state, one at a time. In the current study, the transition from a state to another state is bound by the bigram probabilities described in Section 3. For example, the probability of transition from the state representing $<s>$ to the state representing $/p\text{ɨ}$ is equal to $P(\text{p}|<s>)$.

The model probabilistically generates a phoneme in the source language while it is in a particular state as observation. In the current study, the emission of an observation is bound by the component likelihood in Section 3. For example, the probability of emitting $/b/$ as observation while in the state representing $/p\text{ɨ}$ is equal to $P(b|/p\text{ɨ})$.

The probability of a particular state sequence generating a given observation sequence is calculated by multiplying all the transition probabilities and emission probabilities along the way. As the transition probabilities are equivalent to the component bigram probabilities and the emission probabilities are equivalent to the component likelihood, this is in fact equivalent to multiplying the likelihood and the prior approximated in (5) and (6).

Transition along a number of different state sequences can generate the same observation sequence, but with different probabilities. By running the Viterbi algorithm on the observation sequence, one can identify the state sequence that can generate the given observation sequence with the highest probability (Rabiner 1989: 263–4). In other words, running the Viterbi algorithm will essentially solve the equation in (2).

In this context, the given observation sequence would be the phoneme sequence of the source form. Our goal is to find the most likely sequence of
states from the initial state representing \(<s>\) to the final state representing 
\(</s>\) that can generate the source form. Concatenating the component 
units in the recipient language that the states in the identified sequence 
represent will yield the most likely adapted form predicted by the model.

4.1. Monophones vs. triphones

The hidden Markov model described above emits a single phoneme, or a 
monophone as observation while 
in each state. One can certainly enrich the 
content of the observation in various ways by representing it as a vector 
of multiple feature values instead of a single discrete value; the feature vector 
would consist of a monophone plus values of other features related to the 
monophone.

The feature I consider in this paper is the identity of phones immediately 
adjacent to a given monophone. The idea is that this will capture 
the effect of coarticulation on how listeners in the recipient language perceive 
the individual phones in the source form. The resulting model generates 
triphones as observation instead of monophones.

The difference between the monophone model and the triphone model 
is best illustrated by comparing how they calculate the likelihood. The 
monophone model calculates the likelihood as formulated in (5), which is 
repeated in (7) below for ease of comparison.

\[
P(s \mid r) \approx \prod_{i=1}^{n} P(s_i \mid r_i) = \prod_{i=1}^{n} \frac{C(s_i, r_i)}{C(r_i)} \tag{7}
\]

In (7), the component likelihood \(P(s_i \mid r_i)\) is the emission probability of the 
monophone model, where \(s_i\) is the \(i^{th}\) phoneme in the source form and \(r_i\) 
the corresponding component unit in the adapted form. The triphone model, 
on the other hand, computes the likelihood according to (8).

\[
P(s \mid r) \approx \prod_{i=1}^{n} P(s_{i-1}s_is_{i+1} \mid r_i) = \prod_{i=1}^{n} \frac{C(s_{i-1}s_is_{i+1}, r_i)}{C(r_i)} \tag{8}
\]

In (8), the component likelihood \(P(s_{i-1}s_is_{i+1} \mid r_i)\) is the emission probability of the 
triphone model, where \(s_{i-1}s_is_{i+1}\) is a triphone consisting of the \(i^{th}\) 
phoneme in the source form and its neighboring phonemes. The ‘phonemes’ at the edge, \(s_0\) and \(s_{n+1}\) are word boundary symbols \(<s>\) and \(</s>\).

While a richer representation of observation can in principle lead to a 
more accurate model, it faces the risk of data sparseness. For example, the 
number of triphone types would be the number of monophone types cubed,
in theory. It is possible that some triphones have no token in a given database. In such cases, their emission probability would be zero for every state in the model.

Zero emission probability is problematic because when the model computes the probability of a given observation sequence being generated by a particular state sequence, it multiplies the emission probabilities as well as bigram probabilities. If the emission probability of a triphone were zero for every state in the model, the probability of any state sequence generating the observation sequence including that triphone would be zero. As a result, it would be impossible to identify the most likely state sequence that generated the given observation sequence.

There are several ways to deal with this problem (Jurafsky and Martin 2008: 97-107). For the triphone model in this paper, I use a back-off method such that if every state in the model has zero emission probability for a given triphone, the model backs-off and uses the monophone emission probability as formulated in (9) below. For example, if every state in the triphone model has zero emission probability for the triphone /pzt/, the model uses the emission probability for the monophone /z/ instead.

\[
P(s_{i} \rightarrow s_{i+1} \rightarrow r_{i}) = \begin{cases} 
P(s_{i} \rightarrow r_{i}) & \forall r_{i} : P(s_{i} \rightarrow s_{i+1} \rightarrow r_{i}) = 0 \\ 
P(s_{i} \rightarrow s_{i+1} \rightarrow r_{i}) & \text{Otherwise} \end{cases} \tag{9}
\]

5. Evaluations
Both monophone and triphone models were evaluated in terms of how accurately they predicted the adapted form in Korean for a given English word after being trained on a subset of a Korean loanword database.

5.1. Data
The data used to train and evaluate the model is a subset of the database of Korean loanwords compiled by the National Institute of the Korean Language (NIKL 2008). The original database is a list of 24865 foreign words adopted as Korean loanwords along with information such as the adapted form spelled in Hangul, the domain in which the word is used, the source language, etc. Example entries look like the following:

#1chase#2체이스#3#4편수일반, 용일
#1chashao[叉燒]#2차사오#3중국어#4표준
#1chassis#2섀시#3#4편수일반, 용일, 표준
#1chateaubriand#2샤토브리앙#3프랑스어#4표준
A total of 5812 entries that do not violate the following exclusion criteria were selected from the database:

- The source language is not English.
- The source word is an abbreviation.
- The source pronunciation is not available in a pronunciation dictionary for look-up.
- The adapted form reflects pragmatic knowledge of the word usage (e.g. Adapting ‘Auckland’ as ‘오클랜드 국제 공항’, meaning ‘Auckland international airport’).
- The adapted form is the result of loanword shortening process (e.g. Adapting ‘accelerator’ as ‘악셀’, which is the result of shortening the source word to ‘accel.’).

Each of the selected entries was rewritten as a pair of phonemic transcriptions, one for the original word in English and the other for the Korean loanword. The original source pronunciations were transcribed by looking up the CMU Pronouncing Dictionary, while the adapted forms were phonemically transcribed by applying letter-to-sound rules to their Hangul representation specified in the database. Ultimately, the data consisted of 5812 English-Korean pronunciation pairs.

5.2. Methods

Both the monophone and triphone models were evaluated in terms of prediction accuracy using five-fold cross validation. The dataset of 5812 pronunciation pairs described above was randomly shuffled and split into five sets of similar size. Two of these five sets had 1163 pairs each and the remaining three sets had 1162 pairs each. Each model was evaluated on one of the five sets after its emission probabilities and bigram probabilities were estimated from pairs in the remaining four sets. As there were five sets, this procedure was repeated five times, once for each set.

Prediction accuracy was measured in terms of word accuracy. For a given phonemic transcription of an English word, each model identifies the most likely underlying state sequence by running the Viterbi algorithm. Each state represents a substring of zero or more Korean phonemes. Concatenating the substrings results in a string of Korean phonemes, which was interpreted as the model’s prediction of how the English form will be adapted in Korean. The model’s prediction was deemed correct only if it is exactly identical to the Korean phonemic transcription in the dataset.
5.3. Results

The results are summarized in Table 3. Roughly speaking, the monophone model predicted a little less than half of the test items correctly, while the triphone model predicted a little more than half correctly. Paired t-test showed that the triphone model performed significantly better than the monophone model ($t(4)=18.198, p<0.0001$).

Table 3. Prediction accuracy of the monophone and the triphone models measured in terms of word accuracy using five-fold cross validation.

<table>
<thead>
<tr>
<th></th>
<th>Mean word accuracy</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monophone</td>
<td>48.07%</td>
<td>1.673</td>
</tr>
<tr>
<td>Triphone</td>
<td>52.65%</td>
<td>1.668</td>
</tr>
</tbody>
</table>

5.4. N-best predictions

While the mean word accuracy of the model is more or less than fifty percent, a closer look at the prediction errors reveals that some of them might be alternative answers although they are different from the ones suggested in the original database. For example, consider how some model predictions compare with the “correct” adapted forms in Table 4. There are predictions which might correspond to how some Korean speakers might use the word such as predictions of the triphone model /kʰarɨko/ for ‘cargo’ or /ʌptʰeik/ for ‘uptake’.

Table 4. Example prediction errors compared with the correct adapted forms suggested by NIKL (2008).

<table>
<thead>
<tr>
<th>Source form</th>
<th>Correct</th>
<th>Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>cargo</td>
<td>kʰako</td>
<td>kʰarɨko</td>
</tr>
<tr>
<td>novelty</td>
<td>nopeltʰi</td>
<td>nopeltʰi</td>
</tr>
<tr>
<td>pixel</td>
<td>pʰiksəl</td>
<td>pʰiksəl</td>
</tr>
<tr>
<td>uptake</td>
<td>ʌptʰeik</td>
<td>ʌptʰeik</td>
</tr>
</tbody>
</table>

In other words, there may be alternative forms in which a source form is adapted, but the given database suggests only one of them. As a result, model’s performance may have been underestimated in the results in Section 5.3. One way to deal with this issue is to recruit native speakers of Korean to evaluate the model’s prediction. However, this would be too costly. Another way is to let the model suggest multiple hypotheses and see if the form suggested in the database is one of them.
The Viterbi algorithm can be modified to return the $N$-best list, or a list of $N$ most likely underlying state sequences, rather than the single most likely state sequence (Jelinek 1999: 86-89). So a second experiment was conducted with the same data and methods except for the following. The model was modified to return ten best hypotheses instead of the single best hypothesis. The model’s prediction was deemed correct if the “correct” form was included in the ten best hypotheses.

The results are summarized in Table 5. On average, the ten best hypotheses from the model included the correct answer a little less than seventy percent of the time for the triphone model and a little more than sixty five percent of the time for the monophone model. Paired $t$-test showed that the triphone model performed significantly better than the monophone model ($t(4)=6.120, p=0.0018$).

Table 5. Prediction accuracy of the models measured in terms of the percentage of times the correct answer was one of the ten-best hypotheses from each model.

<table>
<thead>
<tr>
<th></th>
<th>Mean word accuracy</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monophone</td>
<td>65.78%</td>
<td>1.355</td>
</tr>
<tr>
<td>Triphone</td>
<td>69.32%</td>
<td>0.880</td>
</tr>
</tbody>
</table>

6. Conclusion
In sum, I presented a statistical interpretation of the loanword adaptation process which was implemented as a first-order hidden Markov model. As its prediction, the model identifies the most likely sequence of phonemes in the recipient language given the phoneme string representing the source form. The prediction is based on the maximum likelihood estimates of the likelihood of the source form given the adapted form and the prior of the adapted form, where the estimates are derived from a list of pairs of source form and adapted form. Performance of the model was evaluated in terms of how accurately it predicts how English words are adapted in Korean after observing English-Korean pronunciation pairs.

The proposed model is meaningful in two ways. First, the reported performance of the statistical model provides a baseline performance level with which performance of other models can be compared. The proposed model is only one of many potential models of loanword phonology. Furthermore, it is a simple model in that incorporates virtually no linguistic knowledge. Second, it provides the means to evaluate various hypotheses on how the source form is perceived by the listeners of the recipient language. One can represent the source form in different ways and compare
how different representation schemes affect the performance of the model. The current study compared the monophone representation scheme with the triphone representation scheme, but more phonologically interesting representation schemes can be evaluated and compared in the future.

One aspect of the model that is not desirable is that it is supervised. To determine the structure of the hidden Markov model and its parameters, one needs a set of examples of how source forms are adapted in the recipient language. The model must first learn from the examples before it generalizes what it learned to new source forms. This is problematic for two reasons. First, such example data may not be available for some language pairs. In addition, the probabilistic parameters of the model will be more robustly estimated if there are more examples. Even if some example data were available for a given language pair, it may not be large enough. Second, upon hearing a foreign word for the first time, human listeners can come up with some version of the adapted form without recourse to how other foreign words are adopted. An ideal model must mimic the behaviors of the listeners in this regard. Future research should be directed towards making the current model less supervised.

References