Abstract
Variability dependent modeling provides a way of handling the impact of some variability sources in the modeling. In many cases, the variability factor is estimated in a deterministic way, leading to a mere selection of the most adequate model. However, there are always some uncertainty in the estimation of the variability sources which may induce a sub optimal model selection. This paper considers the context of a speaking rate dependent modeling approach, and shows that the uncertainty on the speech segment boundaries, which translates in an uncertainty on the speaking rate estimation, can be handled in the training process and/or in the decoding process. Preliminary results reported here are promising for dealing with variability estimation uncertainty.

Index Terms: speech recognition, pronunciation variants, lexical modeling, speaking rate dependent modeling.

1. Introduction
Many variability sources affect the speech signal received by an automatic speech recognition system [1]. Some are related to the speaker himself, some are related to the environment. Several approaches have been studied for handling one or several of those variability sources. This includes the use of robust acoustic processing and of adaptation techniques. Other approaches have been investigated for introducing the variability estimation in the modeling through hidden variables that take into account for example the so-called speaking mode [2] or the vocal tract length normalization factor [3]. The underlying idea is to reduce the confusability by conditioning the modeling on hidden variables, that are estimated either directly during decoding, or from a previous decoding pass. On the other hand, speech transcription systems still rely on multipass approaches and a set of condition and gender specific models, plus adaptation techniques.

Modeling pronunciation variation is an important topic for automatic speech recognition [4]. Moreover, relevant pronunciation variants are dependent on the speaking style, for example the word final schwa is more frequent in read speech than in spontaneous speech [5]. It has also been observed that speech recognition performance degrades when the degree of spontaneity increases [6] and that the rate of speech also impacts notably on the acoustic realization of the sounds as well as on the pronunciation of the words [1]. Large increase in WER is observed when the speaking rate increases [7]. In [8], a set of parallel rate-specific acoustic and pronunciation models are used to represent slow and fast speech. And, in [9], a speaking rate dependent modeling of the probabilities of the pronunciation variants was proposed.

In this paper we consider the speaking rate dependent modeling of the pronunciation variants as an example of variability dependent modeling, and we focus on the handling of the uncertainty related to the estimation of the variability under consideration, that is, here, on the uncertainty on the speaking rate estimation. This uncertainty comes from the uncertainty on the boundaries of the speech segments used for computing the speaking rate, and is in part due to the HMM approach which do not integrate any constraints and knowledge on the phone boundaries during training.

The outline of the paper is the following. Section 2 recalls the variability dependent modeling framework, as well as the robust procedure used for estimating the probabilities of the pronunciation variants. Section 3 focuses on the handling of the boundary uncertainty in the training and in the decoding processes. Section 4 presents and comments on the recognition experiments that have been conducted on the French ESTER2 speech corpus [10]. Finally conclusions are drawn in section 5.

2. Speaking rate dependent modeling

2.1. Variability Dependent Modeling Framework
In the standard modeling approach, the decoding of an acoustic observation \( x = (x_1, \ldots, x_T) \) consists in finding the most likely sequence \( \hat{w} \) knowing the observation, that is:

\[
\hat{w} = \arg\max_w p(w|x)
= \arg\max_w p(x|w)p(w)
\approx \arg\max_{w,q} p(x|q)p(q|w)p(w)
\]

where \( w = (w_1, \ldots, w_N) \) is a sequence of words, and \( q \) is a sequence of phones corresponding to a pronunciation of the word sequence \( w \).

As proposed in [2], the impact of a given set of variability sources on the modeling, can be taken into account through a variability variable \( v \) estimated from cues \( y \) (note that for speaking rate estimation, the cues \( y \) are usually derived from a first speech decoding pass). This leads to the following decoding equation:

\[
\hat{w} = \arg\max_{w,q} \sum_v p(x|q,v)p(q|w,v)p(w)p(v|y)
\]

where \( v \) represents the variability sources that are taken into account, such as spontaneous speech vs. prepared speech, speaking rate, etc.

In this paper, as a test bed, only the speaking rate variability is considered. Moreover, the impact of the speaking rate variability is analyzed only at the lexicon level (i.e. second term of Equation 2); hence the acoustic and language models are considered independent on the speaking rate in the
remaining part of the paper. Thus, the decoding equation becomes:

\[
\hat{w} = \arg \max_w \sum_y p(x|q)p(q|w, v)p(w)p(y)
\]  

(3)

In a previous set of experiments, reported in [9], the speaking rate variability \( v \) was estimated in a deterministic way, leading to the following decoding equation:

\[
\hat{w} = \arg \max_w p(x|q)p(q|w, v)p(w)
\]

(4)

As each phone sequence \( q \) is the concatenation of a pronunciation variants \( q_i \) of each word \( w_i \) of the word sequence \( w \), the probability of the phone sequence \( q \) is given by:

\[
p(q|w, v) = \prod_i p(q_i|w, v)
\]

(5)

which clearly exhibits the dependency of the probability of a pronunciation variant \( q_i \) of a word \( w_i \) on the whole sequence \( w \) and on the speaking rate variability \( v \). Keeping the dependency on the whole sequence makes it possible to handle phenomena occurring between adjacent words, as for example liaison consonant that can be pronounced only before a word starting with a vowel. Thus, the context \( C_{i+1} \) of the following word \( w_{i+1} \) (i.e. following word \( w_{i+1} \) starting with a vowel or not) is taken into account for modeling the probability of the pronunciation variant \( q_i \) of the word \( w_i \):

\[
p(q_i|w_i, C_{i+1}, v) = p(q_i|w_i, C_{i+1}, v)
\]

(6)

Taking into account the right context dependency of words is typical for the French language because of the liaison phenomenon.

2.2. Estimation of Pronunciation Variant Probabilities

The probabilities of the pronunciation variants are usually estimated from the frequency of occurrences of the pronunciation variants in the training set. However, the training set is limited in size and many words of the lexicon do not occur in the training set. Hence their corresponding pronunciation variant probabilities cannot be directly estimated.

A robust estimation procedure was proposed in [9], based on a set of rules and a MAP estimation. The set of rules corresponds to frequent types of pronunciation variants in French. They refer for example to the presence or absence of the schwa /ə/ in the nth syllable of a word, or in a final position in a word. As a limited number of rules are used, their frequency of occurrences can be rather reliably estimated from the training set, which provide an estimate of their usage probability.

The rules probabilities are then used to compute an a priori probability for the pronunciation variants for which rules apply; otherwise the a priori probabilities are set uniform. Then, using these a priori estimates of the pronunciation variant probabilities and their actual count of occurrences observed in the training set, a Bayesian estimation is made for the probability of each word pronunciation variant. Thanks to the rules used for estimating a priori probabilities, we obtained reasonable probabilities for the pronunciation variants even for words that are not frequent in the training set. On the other hand, the MAP estimation takes benefit of the pronunciation variant frequency counts, especially for frequent words on the training set.

For each word \( w_i \), the speaking rate \( v_i \) is estimated in syllables per second, using a speech segment window spanning a few words \( w_{i-k} \ldots w_i \ldots w_{i+k} \) and centered on the given word \( w_i \). The speech segment may actually be shorter if a speaker change or pauses or hesitations are detected. The speaking rate is quantized in a few bins. Then, for speaking rate dependent modeling, the occurrences in the training set of a pronunciation variant \( q_i \) of a word \( w_i \) are counted according to the context \( C_{i+1} \) defined by the following word and the speaking rate \( v_i \), leading to the counts

\[
\text{count}(q_i; w_i, C_{i+1}, \text{bin}(v_i)), \quad \text{where bin}(v_i) \text{ is the quantized value of } v_i.
\]

3. Handling boundary uncertainty

For a given word \( w_i \), the speaking rate is obtained by dividing the number of syllables in the speech segment centered over the word \( w_i \) by the duration \( \Delta_{\text{alignment}} \) of the segment, as represented in Figure 1. The word boundaries are obtained from speech recognition decoding results.

![Figure 1: Representation of the boundary uncertainty δ for a speech segment \( t_s \ldots t_e \) associated to words \( w_s \ldots w_t \ldots w_e \).](image)

The number of syllables \( nb_{\text{yll}} \) is obtained from the pronunciation variants \( q_s \ldots q_t \ldots q_e \) of the words \( w_s \ldots w_t \ldots w_e \) aligned with the given speech segment. The duration \( \Delta_{\text{alignment}} \) is computed using the indices of the starting frame \( t_s \) of the first word \( w_s \) of the segment, and of the ending frame \( t_e \) of the last word \( w_e \) of the segment. This gives the speaking rate \( v_{i_{\text{alignment}}} \) associated to the word \( w_i \):

\[
v_{i_{\text{alignment}}} = nb_{\text{yll}}/\Delta_{\text{alignment}}
\]

(7)

However, the phone segmentation is error prone, and consequently the word boundaries are not always perfectly located. This induces errors on the computed duration of the segments, and on the estimated speaking rates. The phoneme boundary error is typically of the order of a few frames. The following sub-sections show how the uncertainty on the segment boundaries can be taken into account in the training and/or decoding phases.

3.1. Boundary Uncertainty in Training

In the baseline approach, when counting the number of occurrences of the pronunciation variants on the training set according to the speaking rate, each pronunciation variant \( q_i \) of a word \( w_i \) is associated to the quantized speaking rate \( \text{bin}(v_{i_{\text{alignment}}}) \) of the corresponding speech segment.

Because of the uncertainty on the word boundaries, one can assume that it is likely that the true boundaries are somewhere around the computed boundaries, that is the true beginning is likely to be in \( [t_s - \delta, t_s + \delta] \), and the true ending is likely to be in \( [t_e - \delta, t_e + \delta] \), where \( \delta \) represents the uncertainty on the phone boundaries. That means that the true speaking rate is somewhere between \( v_{i_{\text{slower}}} \) corresponding to the longest duration \( \Delta_{\text{longer}} \) and \( v_{i_{\text{faster}}} \) corresponding to the shortest duration \( \Delta_{\text{shorter}} \).
Although a distribution over the segment durations and corresponding speaking rates could be used to take into account the boundary uncertainty, a simpler approach was used in the current approach. That is, for each word-punctuation occurrence, besides the standard counting associated to the speaking rate \( v_{i,align} \):

\[
\text{count} \left( q_i; w_i, C_{i+1}, \text{bin}(v_{i,align}) \right) = 1 
\]

which counts the number of times the pronunciation variant \( q_i \) of the word \( w_i \) is observed in the right context \( C_{i+1} \) and for the quantized speaking rate \( \text{bin}(v_{i,align}) \), we also take into account the fact that the segment can be shorter or longer due the uncertainty on the boundary positions, and update the corresponding counts accordingly (i.e. the counts associated to a faster or slower speaking rate):

\[
\text{count} \left( q_i; w_i, C_{i+1}, \text{bin}(v_{i,slower}) \right) = \epsilon \delta \\
\text{count} \left( q_i; w_i, C_{i+1}, \text{bin}(v_{i,faster}) \right) = \epsilon \delta 
\]

where \( \epsilon \) is a weight, currently chosen equal to one. Note that this modification of the counts should be refined in order to take into account all the possible speaking rate values between \( v_{i,slower} \) and \( v_{i,faster} \) instead of taking into account only the extreme values as currently done here.

This way of computing the counts of occurrences of the pronunciation variants observed in the training set according to the speaking rate has two effects. First it provides a kind of smoothing between the speaking rate bins. If the computed speaking rate is close to the limit of a speaking rate bin, one of the alternate speaking rate values computed by taking into account the boundary uncertainty is likely to fall in the adjacent bin. Second, taking into account the boundary uncertainty provides more (although somewhat artificial) examples for the extreme (highest & slowest) speaking rate bins which are usually less represented in the data.

### 3.2. Boundary Uncertainty in Decoding

The second way of dealing with the boundary uncertainty is at the decoding level by keeping the summation over the variability variable \( \psi \) in Eq. (3), which leads to computing:

\[
\sum_{v_i} p(q_i|w_i, \text{bin}(v_i)) p(v_i|\psi) 
\]

This summation handles the uncertainty in a probabilistic way through \( p(v_i|\psi) \). In the following experiments, this term is approximated from the computed speech rate \( v_{i,align} \), and the alternate speaking rates \( v_{i,slower} \) and \( v_{i,faster} \) derived by taking into account the boundary uncertainty:

\[
p(q_i|w_i, \text{bin}(v_{i,slower})) p(v_{i,slower}|\psi) \\
+ p(q_i|w_i, \text{bin}(v_{i,align})) p(v_{i,align}|\psi) \\
+ p(q_i|w_i, \text{bin}(v_{i,faster})) p(v_{i,faster}|\psi) 
\]

Although this should be refined, arbitrary values are currently chosen for \( p(v_{i,slower}|\psi) \), \( p(v_{i,align}|\psi) \) and \( p(v_{i,faster}|\psi) \).

4. Experiments

4.1. Experimental Setting

The speech recognition experiments were conducted using the ESTER2 speech corpus [10] which consists of French broadcast news. The data from the African radios were not used here. The counts of the pronunciation variants were obtained on the training set (about 180 hours of signal and 2 M running words). More precisely, the counts were estimated using only the correctly recognized words of the training set. Although some training data (words not correctly recognized) get excluded from the counts, this count estimation was preferred instead of a counting based on a forced alignment, as it avoids having to deal with too many bad (forced) alignments which usually exhibit very short or very long phonetic segments.

The modeling parameters were tuned on a large subset of the development set (about 4h30 of audio signal and 36800 running words). Finally the results are validated on the test set (about 5h50 of audio signal and 63000 running words).

The experiments were conducted using an improved version of the ANTS speech transcription system [11], which is based on the HTK toolkit [12], and the Julius decoder [13][14]. The acoustic models rely on standard cross-word context-dependent modeling, and use a total of about 3000 tied states (shared mixture densities), and 60 Gaussian components per mixture density. The processing includes an HLDA transform on the input features, as well as the use of CMLLR transforms for speaker adaptive training (SAT). The lexicon contains about 63000 words. Because of the optional mute schwa and possible liaisons, there are about 2 pronunciation variants per word, on average. A 4-gram language model is used in the Julius backward pass.

Currently the probabilities of the pronunciation variants are affected by the same fudge factor as the language model. The speaking rates are computed from the speech decoding results of the forward pass of the Julius decoder, and used for conditioning the probabilities of the pronunciation variants used in the backward pass of the decoder. For each word occurring in the results of the forward pass the speaking rate is computed over a few word window (typically 3 to 9 words), centered on this word. Hence, the probabilities of the pronunciation variants are used only in the second (backward) decoding pass of the Julius-decoder.

4.2. Evaluations on Development Set

Table 1 reports a few baseline results on the development set. Using the frequency of occurrences as an approximation of the probabilities of the pronunciation variants provides an improvement over the use of a uniform distribution. Introducing a speaking rate dependent modeling (here a 5 word window is used) provides an extra improvement.

<table>
<thead>
<tr>
<th></th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uniform probability</td>
<td>22.77%</td>
</tr>
<tr>
<td>Frequency of occurrences</td>
<td>22.09%</td>
</tr>
<tr>
<td>Speaking rate dependent modeling</td>
<td>21.95%</td>
</tr>
</tbody>
</table>

The next table reports the impact of taking into account the boundary uncertainty in the training and/or in the decoding processes. Introducing the uncertainty in the training process provides a better estimation of the model parameters and leads to an improvement in the word error rate.
Table 2. Results on the development set taking into account the boundary uncertainty for training and/or decoding.

<table>
<thead>
<tr>
<th>Speaking rate dependent modeling</th>
<th>Uncertainty in training</th>
<th>Uncertainty in decoding</th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No</td>
<td>No</td>
<td>21.95%</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>No</td>
<td>21.83%</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>Yes</td>
<td>21.95%</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>Yes</td>
<td>21.85%</td>
</tr>
</tbody>
</table>

Finally a few experiments were conducted to analyze the impact of the size of the window used for computing the speaking rate for each word. The results reported in Table 3. show that using a somewhat larger window, 9 words, provides better results than the 5-word window used above.

Table 3. Impact of the size of the window used for computing the speaking rate for each word.

<table>
<thead>
<tr>
<th>Windows</th>
<th>3 words</th>
<th>5 words</th>
<th>7 words</th>
<th>9 words</th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>21.99%</td>
<td>21.95%</td>
<td>21.96%</td>
<td>21.88%</td>
<td></td>
</tr>
</tbody>
</table>

In the previous experiments, the boundary uncertainty value δ was set to 30 ms. Some experiments carried on using a value of 50 ms did not provide better results. However, no optimization of this parameter has been carried on yet.

4.3. Validation on Test Set

To validate the approach proposed in this paper for handling boundary uncertainty in a speaking rate dependent modeling, an evaluation has been conducted on the test set. The results obtained using a 9-word window for estimating the speaking rate, and a 30 ms uncertainty for the boundaries, are reported in Table 4.

Table 4. Recognition results on test set.

<table>
<thead>
<tr>
<th>Uniform probability</th>
<th>Uncertainty in training</th>
<th>Uncertainty in decoding</th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>---</td>
<td>---</td>
<td>22.94%</td>
</tr>
<tr>
<td>Frequency occurrences</td>
<td>---</td>
<td>---</td>
<td>22.63%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Speaking rate dependent modeling</th>
<th>Uncertainty in training</th>
<th>Uncertainty in decoding</th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No</td>
<td>No</td>
<td>22.39%</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>No</td>
<td>22.31%</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>Yes</td>
<td>22.35%</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>Yes</td>
<td>22.30%</td>
</tr>
</tbody>
</table>

The results show that the detailed modeling of the probabilities of the pronunciation variants based on a speaking rate dependent modeling provides better results than the simple usage of the frequency of occurrences of the pronunciation as an estimate of the probabilities.

Moreover the results show that taking into account the boundary uncertainty either during training or during decoding, or both, provides a slight improvement over the rough speaking rate dependent modeling approach.

5. Conclusions

In this paper the speaking rate dependent modeling of the probabilities of the pronunciation variants was used as a test bed for investigating the handling of uncertainty on the variability estimation. Here the uncertainty comes from errors in the location of the phone boundaries used to measure the speech segment durations for computing the speaking rates.

The uncertainty on the variability estimation, here the speaking rate, was first handled in the training process. This had two effects. On one hand, this provides a smoothing of the estimated counts of occurrences between quantized speaking rate bins. On the other hand, this increased the amount of examples falling in extreme bins (slowest and fastest speaking rate bins), which are often very poorly represented in the data. However, this also generally corresponds to the variability values that have the largest impact on the modeling and on the speech recognition performance. The second approach consists in keeping and handling the uncertainty in the decoding process itself. Of course both approaches can be combined.

Experiments conducted on the French ESTER2 speech corpus, using a speaking rate dependent modeling of the probabilities of the pronunciation variants showed a slight improvement in the WER when the uncertainty was taken into account either in the training process or in the decoding process. This is an interesting result that leads to extending, in the future, this approach to other kinds of variability sources.

6. References