UNIVERSIFIED HMM ADAPTATION BASED ON SPEECH-SILENCE DISCRIMINATION

Ilija Zeljković, Shrikanth Narayanan and Alezandros Potamianos

AT&T Labs – Research, 180 Park Avenue, P.O. Box 971, Florham Park NJ 07932
e-mail: {iliija, shri, potam}@research.att.com

ABSTRACT

An unsupervised, sentence-level, discriminative, HMM adaptation algorithm based on silence-speech classification is presented. Silence and speech regions are determined either using an endpointer or using the segmentation obtained from the recognizer in a first pass. A unsupervised discriminative training procedure using the gradient descent algorithm, with N-best competing strings with word insertions is then used to improve the discrimination between silence and speech. Experiments on connected digits show about 40-80% reduction in insertion errors, a small amount of reduction in substitution errors, and a negligible effect on deletion errors. In addition, experiments on noisy speech showed about 70% word error rate reduction, thus demonstrating the robustness of the proposed adaptation technique.

1. INTRODUCTION

Acoustic mismatch between training and testing conditions results in significant performance degradation in hidden Markov model (HMM)-based speech recognizers. Careful inspection of the recognition errors shows that word insertion and substitution errors often occur as a result of poor recognition scores for acoustic segments with low energy. Low-energy portions of the speech signal tend to be highly confusible with silence, especially when channel and noise mismatch exists between the speech signal and the acoustic models. Various blind deconvolution and bias removal schemes (for e.g., cepstral mean normalization [1]) exist that address this problem of mismatch at the utterance level. This paper focuses on reducing the utterance-model mismatch at those segments of the speech signal where the acoustic characteristics of the background and the speech signal, typically unvoiced portions, are similar.

Our goal in this paper is to adapt the parameters of the acoustic model, in an unsupervised mode, during the recognition process in order to improve discrimination between the background model and the speech models. One (“instantaneous adaptation”) or more (“long-term adaptation”) utterances can be used for adaptation. The main idea is to increase the separation between the correct string and competing string candidates. Although the correct string is not known during the recognition process, this study exploits the fact that the silence regions of the sentence can be determined with greater accuracy and efficiency and these regions can be used for discriminative training.

The organization of this paper is as follows. First we present an overview of the relevant literature followed by the description of our proposed speech-silence discrimination adaptation algorithm. In Section 2, implementation issues are discussed. Specifically, novel ways of producing competing recognition hypothesis for the adaptation algorithm are proposed. Finally, in Section 3, we apply the algorithm to a connected digit recognition task and show digit error rate reduction up to 70-80%.

1.1. PREVIOUS WORK

A significant part of the speech recognition literature deals with problems caused to real-world recognition systems by noise, distortion or variability in the speech waveform. Various algorithms such as cepstral mean normalization, maximum likelihood (ML) cepstrum bias normalization [1], ML frequency warping [2], and ML linear regression [3] have been proposed to deal with these problems. Apart from these transformation-based techniques that produce good results with a limited amount of adaptation data, the acoustic models can be retrained using Bayesian adaptation. Bayesian adaptation typically requires a large amount of adaptation data; algorithms have been proposed for updating groups of HMM parameters or for smoothing the re-estimated parameter values, for e.g., vector field smoothing, classification tree or state-based clustering of distribu-
tions [4]. Parallel model combination (PMC) has been also used to combat both additive noise distortion and multiplicative (channel) distortion [5].

Typically, these algorithms perform well for simulated data, i.e., when additive or multiplicative distortion is added to the speech signal in the laboratory but not equally well in field trials where a multitude of sources with time-varying characteristics can distort the speech signal. In many cases, very little data are available for adaptation. Further, the adaptation data might not be transcribed. For example, discriminative training of HMMs, which helps improve recognition accuracy [8, 6, 7], assumes that the linguistic context of the utterance is known. Unsupervised adaptation using very few utterances is a difficult problem because there are no guarantees that the adapted parameters will converge toward global optimum values. Utterance verification algorithms can be used to make unsupervised adaptation more robust.

In this paper, we focus on the problem of adapting the speech and silence acoustic models to improve our ability to discriminate between speech and silence regions. In the next section, we propose novel ways of producing competing recognition hypotheses. Unsupervised discriminative training with generalized probabilistic descent algorithm (GPD) is used to adapt the speech and silence acoustic models [9].

2. SPEECH-SILENCE ADAPTATION ALGORITHM

The short-term and long-term spectral characteristics of the background part of speech utterances collected through the public switched telephone network are highly variable. Although HMM-based recognizers often confuse the background with valid speech segments, thereby producing insertion or deletion errors, portions of the background (silence) regions can be identified with fairly high accuracy using simpler but more robust techniques using features such as the short-time energy and the short-time zero-crossings. The high degree of certainty in determining silence regions and the use of only these regions to adapt both speech and silence models makes the speech-silence adaptation algorithm both accurate and efficient.

The adaptation algorithm consists of the following four steps:

1. Split the input utterance into speech and silence regions. If more than one background HMM is used, do optimal (Viterbi) decoding of silence regions using the background HMMs.
2. Generate competing strings by aligning the states of speech HMM to silence regions (i.e., artificially encourage or force insertions).
3. Enhance separation between HMMs in correct and competing strings using a discriminative training algorithm.
4. Perform optimal decoding (recognition) on the whole utterance using the newly adapted HMMs and any prescribed grammar.

There are many ways to implement steps 1, 2 and 3 of the algorithm proposed above. Next, we discuss a simple and efficient implementation of the adaptation algorithm.

2.1. IMPLEMENTATION ISSUES

Speech-silence segmentation (step 1) may be obtained by a simple preprocessing step before the recognition process begins. In the current implementation the silence-speech segmentation is performed by the recognizer in the first pass using the initial HMMs, and a grammar, with no insertion penalties. This is assumed to be the “correct string”. Competing strings (step 2) are produced in two alternative ways:

(a) Acoustically-driven insertion: A negative insertion penalty (insertion incentive) is used to decode four best competing strings (encouraged internal insertion).

(b) Blind external insertion: Eleven competing strings (for digit recognition test) are generated: each digit is added before and after the initially recognized string, generating one competing string (forced external insertion).

The discriminative training [7, 6] (step 3) is performed by using the minimum string-error training algorithm using $N$ competing string models [9]. A brief description of the discriminative training algorithm using GPD is given in the next section (Sec. 2.2). Finally the second-pass recognition is performed with the adapted models using the Viterbi decoding algorithm (step 4).

2.2. DISCRIMINATIVE TRAINING

The goal of the discriminative model training algorithm [7, 6] is to find a model set that optimally distinguishes between observation sequences corresponding to correct class models and those of $N$
competing class models by maximizing the mutual
information between the observation sequence \( O \)
and the words or strings of that class (represented
by a parametrized HMM, \( \Lambda \)). The misclassifica-
tion measure

\[
d(O, \Lambda) = -g(O, S_\nu, \Lambda) + \log \left\{ \frac{1}{N-1} \sum_{S_k \not= S_\nu} e^{g(O, S_k, \Lambda)} \right\}^{\frac{1}{K}}
\]

uses the discriminant function

\[
g(O, S_k, \Lambda) = \log f(O, \Theta S_k, S_k | \Lambda)
\]

which is defined in terms of the loglikelihood score
function \( f \) on the optimal state sequence \( \Theta S_k \)
(given the model set \( \Lambda \)) for the \( k^{th} \) best string, \( S_k \). The
discriminant function for the transcribed training
string \( S_\nu \) is \( g(O, S_\nu, \Lambda) \). The model loss function
for string error rate minimization, \( \ell(O, \Lambda) = 1/(1 + \exp(-\gamma d(O, \Lambda))) \), where \( \gamma \) is a positive constant, is solved using gradient descent algorithm
[8, 9].

As will be shown in the next section, the \( N \)
strings compete among each other directly from the
acoustics or externally by some simple blind string
appending scheme.

3. RECOGNITION EXPERIMENTS

Speech units (words and subwords) as well as back-
ground silence are modeled by first order, left-to-
right HMMs with observation densities. The observation vector consists of 39 features: 12 LPC derived cepstral coefficients, dyna-
mically normalized energy, as well as their first and second derivatives. Eleven digits, including "oh" and "zero", were used in the evaluation task. Each digit was modeled with either 20 or 15 state
HMMs, with 16 Gaussian mixtures. Speech background (silence) is modeled with a single state, 128
Gaussian mixture HMM. The HMMs were trained using data extracted from speech data collected
over the telephone network (16089 digit strings).

In the recognition process, the sequence of ob-
servation vectors from an unknown speech utter-
ance are matched against a set of stored hidden
Markov models representing speech units. A search network is generated by a finite state gram-
mar that describes the set of valid strings. The network search algorithm returns the single most
likely sequence of speech units. The search pro-
cedure is a Dynamic Programming (DP) algorithm
(Viterbi decoding) where the goal is to find a
valid state sequence with the highest accumulated
state log-likelihood.

Experimental Results: The algorithm was
tested on speech data collected from two AT&T
service trials. Trial I data, consisting of 10768
16-digit strings, represented matched training and
testing conditions. On the other hand, no data
from Trial II were represented in training. More-
over, Trial II data consist only of single digits (a
total of 2159 utterances). It should be pointed out
that isolated digits represented only a small por-
tion of the training database. In addition, a test
set with noisy speech data (601 digit strings spoken
over the telephone network), marked by human
listeners, was used to verify the robustness of the
proposed algorithm.

Tables 1 and 2 summarize the recognition results for various testing conditions: Results are compared for the two methods of competing string generation (N-best competing strings by
acoustically-driven insertion using insertion incen-
tives and blind external insertion by forced initial and final digit appending), with each case repeated with and without resetting the models to the baseline class for each new string input. The baseline results correspond to no model adaptation.

Under reasonably matched training and testing
conditions, we observe that insertion errors are
reduced in all test cases when adaptation is used.
The best results are obtained for the case that uses
competing strings generated through insertion incen-
tives. Moreover, as expected, long-term adap-
tation (using all available utterances for adapta-
tion) performs better than instantaneous adapta-
tion (i.e., a single utterance is used to adapt the
HMMs). On the other hand, although the blind
insertion method has a similar effect on insertion
errors, it is accompanied by increased substitu-
tion and deletion errors, particularly in the long-
term adaptation case, suggesting divergence in the
adapted models with increasing adaptation data.

The unusually high number of insertion errors
in the baseline results for Trial II data is attrib-
tuted to the structural mismatch between the
training data and this particular test set which is
composed entirely of isolated digits. Instantan-
eous adaptation gives about 36-38% improvement
in word error rates for both methods of compet-
ing string generation. For long-term adaptation,
however, the blind insertion method of competing
string generation yields poorer performance than
the baseline while the acoustically-driven inser-
tion method yields more than 80% improvement
in word error rate. A closer analysis of the re-
results shows that although there is improvement in
insertion errors (which is indeed the objective of
our proposed algorithm), there is significant in-
crease in substitution errors for the blind inser-
tion method. This result further supports our earlier
remark that model divergence (instability) with in-
creasing adaptation data is a potential pitfall when
blind insertion is used for competing string gener-
Table 1: Recognition performance on Trial I test data (N = 10768 strings) with HMM adaptation: matched training and testing conditions. Results are shown for instantaneous and long-term adaptation and for the two methods of competing string generation: acoustically-driven by insertion incentives and blind insertion by word appending.

<table>
<thead>
<tr>
<th>Competing string generation</th>
<th>Adaptation mode</th>
<th>Total</th>
<th>Sub</th>
<th>Del</th>
<th>Ins</th>
</tr>
</thead>
<tbody>
<tr>
<td>None (baseline)</td>
<td>N/A</td>
<td>1.25</td>
<td>0.8</td>
<td>0.1</td>
<td>0.4</td>
</tr>
<tr>
<td>acous. driven ins</td>
<td>long-term</td>
<td>1.05</td>
<td>0.8</td>
<td>0.1</td>
<td>0.2</td>
</tr>
<tr>
<td>blind ins</td>
<td>long-term</td>
<td>1.23</td>
<td>0.9</td>
<td>0.2</td>
<td>0.2</td>
</tr>
<tr>
<td>acous. driven ins</td>
<td>instant.</td>
<td>1.16</td>
<td>0.7</td>
<td>0.1</td>
<td>0.3</td>
</tr>
<tr>
<td>blind ins</td>
<td>instant.</td>
<td>1.17</td>
<td>0.7</td>
<td>0.1</td>
<td>0.3</td>
</tr>
</tbody>
</table>

Table 2: Recognition performance on Trial II test data (N = 2159) with HMM adaptation: mismatched testing and training conditions. Other testing conditions similar to those for Table 1.

<table>
<thead>
<tr>
<th>Recognition details</th>
<th>Word Error</th>
<th>Sub</th>
<th>Del</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>15.9</td>
<td>13.8</td>
<td>2.2</td>
</tr>
<tr>
<td>Blind ins</td>
<td>4.9</td>
<td>2.2</td>
<td>0.9</td>
</tr>
</tbody>
</table>

Table 3: Recognition performance on digit strings (N = 601) with appreciable background noise (as marked by human listeners) with instantaneous HMM adaptation.

4. SUMMARY

A novel HMM adaptation method based on speech-silence discrimination was presented. In summary, the main contributions of this work are:

- The exclusive use of signal portions declared by the algorithm as silence segments (i.e., unsupervised modality) for adapting both silence and some/all speech models in a way that results in improved speech-silence discrimination in the new model set.
- Automatic competing string generation by providing insertion incentives, inserting words that are naturally prone to acoustic confusion with background.
- Unsupervised adaptation using the gradient descent algorithm that assures convergence.

Results show that competing strings directly provided by the recognizer by employing insertion incentives give the most useful set of data for speech-silence discrimination, and yields the best overall error rate improvements even under mismatched training and testing conditions. Dramatic improvements in insertion errors were obtained for telephone speech data with high background noise.

5. REFERENCES


