Cross-System Adaptation and Combination for Continuous Speech Recognition: The Influence of Phoneme Set and Acoustic Front-End

Sebastian Stükere, Christian Fügen1, Susanne Burger2, Matthias Wölfel1

1Institut für Theoretische Informatik, Universität Karlsruhe (TH), Karlsruhe, Germany
2Interactive Systems Laboratories, Carnegie Mellon University, Pittsburgh, USA
{stueker|fuegen|wolfel}@ira.uka.de, sburger@cs.cmu.edu

Abstract

Cross-system adaptation and system combination methods, such as ROVER and confusion network combination, are known to lower the word error rate of speech recognition systems. They require the training of systems that are reasonably close in performance but at the same time produce output that differs in its errors. This provides complementary information which leads to performance improvements. In this paper we demonstrate the gains we have seen with cross-system adaptation and system combination on the English EPPS and RT0-05S lecture meeting task. We obtained the necessary varying systems by using different acoustic front-ends and phoneme sets on which our models are based. In a set of contrastive experiments we show the influence that the exchange of the components has on adaptation and system combination.

Index Terms: automatic speech recognition, system combination, cross adaptation, EPPS, RT-05S.

1. Introduction

In state-of-the-art speech recognition systems it is common practice to use multi-pass systems with adaptation of the acoustic model in-between passes. The adaptation aims at better fitting the system to the speakers and/or acoustic environments found in the test data. It is usually performed on a by-speaker basis, obtained either from manual speaker labels or automatic clustering methods. Common adaptation methods try to transform either the models used in a system or the features to which the models are applied.

Three adaptation methods that can be found in many state-of-the-art systems are Maximum Likelihood Linear Regression (MLLR) [1], a model transformation, Vocal Tract Length Normalization (VTLN) [2] and feature-space constrained MLLR (fMLLR) [3], two feature-transformation methods. Adaptation is performed in an unsupervised manner, such that the error-prone hypotheses obtained from the previous decoding pass are taken as the necessary reference for adaptation. Generally, the word error rates of the hypotheses obtained from the adapted systems are lower than those for hypotheses on which the adaptation was performed. This sequences of adaptation and decoding make it possible to incrementally improve the performance of the recognition system. Unfortunately, this loop of adaptation and decoding does not always lead to significant improvements. Often, after two or three stages of adapting a system on its own output, no more gains can be obtained. This problem can be overcome by adapting a system S2 on the output of a different system S1, a process commonly referred to as cross-system adaptation. It is believed that the gains from cross-system adaption come from the fact that S1 makes different errors than S2. S2 thus gets complementary information that it could not gain from its own output. It is also possible to utilize the complementary information contained in hypotheses from different recognition systems by using system output combination methods, such as ROVER [4] and confusion network combination (CNC) [5].

For both methods it is necessary to build multiple systems that are reasonably close in performance to each other, but which produce hypotheses with complementary knowledge. We report on our experiences with adapting across systems which vary in phoneme set and acoustic front-end, and the combination of outputs using CNC. We report and compare results on the English European Parliamentary Speeches Task [6] and the Lecture Room task of the NIST Rich Transcription 2005 Spring Meeting Recognition Evaluation (RT-05S). The next section describes previous related work and how our work differs from it. Section 3 describes and compares the two phoneme sets used for the experiments, while Section 4 introduces the acoustic front-ends applied in our experiments. Section 5 provides the results of the experiments.

2. Related Work

For their NIST 2004 Fall Mandarin Broadcast News evaluation system [7], Yu et al. used two different kinds of models; one set based on phonemes, the other based on initial-final semi-syllables. The two sets of models were used for cross-adaptation and for system combination. In our work, all
sets of models are based on phonemes, since for English
syllable-based models have generally been found not to be
competitive with phoneme-based models.

In [8], Stolcke et al. used two different kinds of front-
ends, one MFCC and one PLP based, for cross-adaptation
and system combination via confusion networks. They did
not change their phoneme set for the different systems,
while we varied the phoneme set for the models. We fur-
ther used an MVDR front-end instead of a PLP front-end,
since we found it to be superior to PLP in many tasks.

Lamel and Gauvain experimented in [9] with different,
either reduced or extended, versions of the same phoneme
set, used them in a cross-adaptive way and combined the
system results with ROVER. Though the performances of
the different phoneme sets were basically the same, ROVER
gave a significant improvement. The front-ends remained
unchanged. Also, the dictionaries for the different phoneme
sets were essentially created from the same base dictionary,
while in our experiments the dictionaries for the phoneme
sets were derived from differing base dictionaries, and mis-
ing pronunciations were created with differing tools.

3. Phoneme Sets and Dictionaries

3.1. Phoneme Sets

(We use the term phoneme rather than phone because even
though the described sets include a few allophones, they
are working on phoneme level.) We experimented with the
CMU dictionary - CMUDICT, and LDC’s Call Home dic-
tionary - Pronlex. Our version of CMUDICT consists of 45
phonemes and allophones and our version of Pronlex con-
tains 44 phonemes and allophones. Despite using a slightly
different approach regarding the symbols used to represent
the phonemes, the inventories are the same for the five diph-
thongs or vowel-glide sequences {eoYOW} {EY OW AY
OY AW}, nine fricatives {szSZfTDh} {S Z SH ZH F V
TH DH HH}, two affricates {CJ} {CH JH}, six plosives
{pbtdkg} {P B T D K G} and three nasals {mnG} {M N
NG}. Both systems contain the seven vowels {iIE@acU}
{IY IH EH AE AA AO UH}. We used the extended Pronlex
set to include {A u} which map to the already existent {AH
UW} in CMUDICT (e.g. the vowels in “two” and “hut”).
There are four approximants {lrwy} {L R W Y} in both
systems. Pronlex additionally allows for an allophone of the
voiced velar approximant {w}: a voiced velar approximant
with initial velar friction noted as {H} (it sounds like a /h/
followed by /w/). CMUDICT only uses {W} which denotes
the version with more initial friction and the version without
friction. The systems also differ in the number of reduced or
centralized vowels: CMUDICT uses {IX} for centralized /u/
(for example, in the last syllable of “laughing”) but also uses
a symbol for a short lowered closed front vowel: {IH}. In
Pronlex both the lowered close front vowel and the central-
ized version of it are labeled with {I}. Both systems provide
symbols for the mid central vowel {x} {AX} (e.g. the final
sound in “Maria”) and the open central vowel {R} {ER}
(e.g. in “hurt” or the final sound in “father”). However, our
extended version of CMUDICT differentiates between the
the final sounds of “father”: {ER} and “answer”: {AXR}.

3.2. Dictionaries

The necessary pronunciation dictionaries for training and
testing were created in different ways for the two differ-
ent phoneme sets. In the case of the Pronlex phoneme
set, the initial version of all lexicons was a merger of the
Callhome_english_lexicon_97061 and the LIMSI SI-284
training dictionary. Frequent missing words were added
by hand, all other words were generated with the help
of William Fisher’s grapheme-to-phoneme tool available
through NIST [10]. For the CMUDICT phoneme set, we
used the dictionary from the ISL Meeting Transcription
System [11] as a base dictionary and created missing pro-
nunciations using Festival [12].

4. Acoustic Front-Ends

In our experiments we used four different kinds of acoustic
front-ends: MFCC-I, MFCC-II, MVDR-I, and MVDR-
II. Two are based on the traditional Mel-frequency Cep-
stral Coefficients (MFCC) and two are based on the warped
minimum variance distortion less response (MVDR). The
second front-end replaces the Fourier transformation by a
warped MVDR spectral envelope [13], which is a time do-
main technique to estimate an all-pole model using a warped
short time frequency axis such as the Mel scale. The use of
the MVDR eliminates the overemphasizing of harmonic peaks
typically seen in medium and high pitched voiced speech
when spectral estimation is based on linear prediction.

For training, both front-ends have provided features ev-
ery 10 ms. During adaptation and decoding this was some-
times changed to 8 ms. In training and decoding, the fea-
tures were obtained either by the Fourier transformation fol-
lowed by a Mel-filterbank or the warped MVDR spectral
envelope.

We used a model order of 80 for the MVDR-I front-end.
The resulting 129 spectral coefficients were then reduced to
30 with a linear filterbank. Since the warped MVDR already
provides the properties of the Mel-filterbank, namely war-
ping to the Mel-frequency and smoothing, a filterbank has
not been used for the MVDR-II front-end and the model
order was just 22. The advantage of this approach is an
increase in resolution in low frequency regions. This can-
ot be attained with traditionally used Mel-filterbanks and
unequal modeling of spectral peaks and valleys used to im-
prove noise robustness, due to the fact that noise is mainly
present in low energy regions.
Table 1: Result Overview of the cross adaptation experiments for the EPPS task. Adaptation is performed on the CNC output from the second stage of the adaptation scheme.

<table>
<thead>
<tr>
<th>Phoneme Set</th>
<th>Acoustic Front-End</th>
<th>8ms</th>
<th>10ms</th>
<th>8ms</th>
<th>10ms</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MFCC-II</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>MVDR-II</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CMU</td>
<td>13.7%</td>
<td>14.0%</td>
<td>13.8%</td>
<td>13.7%</td>
<td></td>
</tr>
<tr>
<td>Pronlex</td>
<td>14.6%</td>
<td>14.6%</td>
<td>14.6%</td>
<td>15.0%</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Results Overview of cross system adaptation on RT-05S-eval. The Pronlex system uses an MVDR front-end and was adapted on the CNC output of the 3rd pass. CNC with the Pronlex system was done by using also the lattices of the 3rd pass systems.

<table>
<thead>
<tr>
<th>pass</th>
<th>CMU-I</th>
<th>CMU-II</th>
<th>PRON-MVDR</th>
<th>CNC</th>
</tr>
</thead>
<tbody>
<tr>
<td>3rd</td>
<td>24.9%</td>
<td>25.4%</td>
<td>23.9%</td>
<td></td>
</tr>
<tr>
<td>4th</td>
<td>25.0%</td>
<td>24.8%</td>
<td>23.8%</td>
<td></td>
</tr>
<tr>
<td>4th</td>
<td></td>
<td></td>
<td>24.6%</td>
<td>23.2%</td>
</tr>
</tbody>
</table>

For all front-ends, VTLN was applied either in the linear domain for MFCC-I and MFCC-II, or in the warped frequency domain for MVDR-I and MVDR-II. The MFCC uses 13 cepstral coefficients while for the MVDR the number of cepstral coefficients has been increased to 15 (EPPS) or 20 (RT-05S). The mean and variance of the cepstral coefficients were normalized on a per-utterance basis. In the case of MFCC-I, MVDR-I, and MVDR-II, seven adjacent frames were combined into one single feature vector. For MFCC-II the cepstral coefficients were combined with normalized signal energy, approximations of the first and second derivative, and zero crossing rate. For MFCC-I, MVDR-I, and MVDR-II, the resulting feature vectors were then reduced to 42 dimensions using linear discriminant analysis (LDA). LDA was applied to MFCC-II without dimension reduction.

5. Experiments

All experiments were performed with the help of the Janus Recognition Toolkit (JRTk) featuring the IBIS single pass decoder [14]. The systems described below have, at least in part, been used for the Spring 2006 TC-STAR EPPS evaluation [15] and the RT-06S Lecture Task [16].

5.1. European Parliamentary Speeches

The European Parliamentary Speeches Task (EPPS) focuses on transcribing speeches given in the European Parliament. The word error rates in the experiments reported below were measured on the official 2006 development set, which consists of three hours of speech from 41 politicians. The acoustic models were trained on the official EPPS training data which consists of about 100 hours of transcribed speech from politicians and interpreters. Before starting the cross-system adaptation experiments we first ran two adaptation stages which used only CNC for system combination. In the first stage we performed two decodings with speaker independent systems. Both use the Pronlex phoneme set based dictionary, but one utilizes the MVDR-II front-end, while the other uses the MFCC-I front-end, both with a frame shift of 10 ms. The two outputs are then combined using CNC. Then, in the second stage, three acoustic mod-
tributions over 3,000 models, also with a maximum of 64 Gaussians per model. All systems were trained with either ML-SAT or FSA-SAT and use the same vocabulary and language models for decoding.

Table 2 shows a part of our RT-06S evaluation system. As can be seen, the cross-system adaptation of the CMU-I and CMU-II system leads to no further improvements. Even though the CMU-II system improves in the fourth pass by an absolute value of 0.6%, the confusion network combination of the lattices of the same pass only changed by 0.1%. But if we adapt the PRON-MVDR system on the CNC output of the third pass and do a confusion network combination on the lattices from the CMU-I and CMU-II system of the third pass and the Pronlex system in the fourth pass, we can improve the CNC output by an absolute value of 0.7%.

6. Conclusions

In decoding set-ups in which the models of the system are incrementally adapted on the output of previous decoding passes, the models are often saturated after two or three iterations of adaptation. Further adaptation steps on the output from the same system yield no more significant gains. However, when using the output of systems that differ in some components, it is possible to obtain further gains due to complementary knowledge. In our experiments we have shown how systems with different phoneme sets and acoustic front-ends can be used in a cross-system adaptation scheme in order to get higher gains out of adaptation. Further we have shown how the outputs from the different systems can be combined using confusion network combination, leading to further reductions in word error rate.

7. Acknowledgments

This work has been funded in part by the European Union under the integrated projects TC-Star - Technology and Corpora for Speech to Speech Translation - (IST-2002-FP6-506738, http://www.tc-star.org) and CHIL - Computers in the Human Interaction Loop (IST-506909, http://chil.server.de). The authors would like to thank Joseph P. Fridy for his help in proof-reading the paper.

8. References