ABSTRACT

In this paper, hybrid HMM/ANN systems are used to model context dependent phones. In order to reduce the number of parameters as well as to better catch the dynamics of the phonetic segments, we combine (context dependent) diphone models with context independent phone models. Transitions from phone to phone are modeled as generalized context dependent distributions while phonetic units are context independent models trained on the less coarticulated middle part of each phone. Words are thus modeled as a sequence of probability distributions alternatively representing the middle part of the phonemes and the transitions from phone to phone. A single neural network is used to estimate both context independent phone probabilities and generalized context dependent diphone (phone to phone transition) probabilities. Resulting systems are compared to classical context independent phone-based HMM/ANN systems with the same number of parameters. The Phonebook isolated word database has been used for training the systems. Testing is done on small (75 words), medium (600 words) and large (8000 words) lexicons. Test words were not present in the training vocabulary.

1. INTRODUCTION

Task (or vocabulary) independent training remains an important issue in current automatic speech recognition (ASR) systems. It is indeed well known that ASR performance is always significantly lower for lexicon words that were not observed in the training data. Most state-of-the-art ASR systems address this problem by using generalized context dependent phoneme models [5].

Supported by a F.R.I.A. grant (Fonds pour la Formation à la Recherche dans l’Industrie et dans l’Agriculture).

2. APPROACH

As shown in Figure 2, we used a single ANN to estimate both context independent phoneme probabilities and generalized context dependent diphone probabilities.
2.1. Phone to phone transition clustering

Two strategies have been used to cluster the whole set of transitions into a limited number of generalized classes.

The first one was to define transition classes based on the broad phonetic classes of the left phone and of the right phone. As opposed to transitions (from the complete set of transitions), which are defined by the left and right phones, the generalized transitions were characterized by the left and right phonetic classes. It was indeed assumed that left or right phones from the same phonetic class have similar effects on the acoustic realization of the transition. We have considered two criterions to define the phonetic classes:

- silence + 9 classes based on the broad phoneme category. Table 1 describes the classes defined around the TIMIT phoneme subset that will be used in our experiments.

- silence + 8 classes based on the place of articulation. These classes are described in Table 2.

The second strategy was based on automatic data-driven clustering of all the phone to phone transitions. In this approach, we identified states for which pdf’s could be tied with a minimal loss of the system modeling capability, avoiding introducing a priori (and usually inaccurate) knowledge in the system. The pdf’s of a given state were modeled using single gaussian distributions and a K-means algorithm was used to cluster the gaussians. The distance measure between two gaussians was defined by

\[ d(i, j) = \frac{1}{V} \sum_{k=0}^{V} \frac{(\mu_k - \mu_{jk})^2}{\sigma_k^2 + \sigma_{jk}^2} \]

where \( \mu_k \) is the mean vector and \( \sigma_k \) is the standard deviation vector of gaussian \( i \).

2.2. Discussion

The context dependent hybrid HMM/ANN model is expected to lead to better recognition performance in the case of training independent tasks. Indeed, the training data always contain the phonemic units in a limited number of left and right contexts. Therefore, in standard recognizers (and standard reference tasks, i.e. vocabulary dependent tasks), it is possible to make use of the contextual information to improve performance, e.g., by using context dependent phone model. Even when using context independent phone models (which is often the case with hybrid HMM/ANN systems that are however known to yield comparable – although still somewhat lower – performance compared to context-dependent CDHMMs models), the phone models will implicitly capture some contextual information. However, if
the application (or test) vocabulary is different from
the training vocabulary, using that phonemic con- 
textual information could result in a loss of per- 
formance since the trained models are no longer really
appropriate. The context dependent hybrid system
attempts to limit this effect by tying the distribu- 	ions of context dependent models. In this way, it 

be expected that all major contextual effects will 
be captured by those tied “transition” states 
while the context independent phoneme models will
focus on the actual “steady-state” section of each
phonemic segment.

3. DATABASE

All the experiments reported in this paper have been 
carried out on the Phonebook [6] database. This is 
a phonetically-rich isolated word telephone-speech 
database. Phonebook consists of more than 92,000
utterances and almost 8,000 different words, with an
average of 11 talkers for each word. Each speaker of
a demographically-representative set of over 1,300
native speakers of American English made a single
telephone call and read 75 words.

The database contains 106 word lists, each com- 
piled of 75 or 76 words that have been pronounced
by a few (typically around 11) speakers. The speak- 
ers and words are different for each word list. The
word lists are labeled as \( l_1 \) with

\[ l_1 \in \{a, b, c, d, e\} \]

and \( l_2 \in \{a, b, c, d, e, f, \ldots, x, y, z\} \)
except if \( l_1 = e \), in which case \( l_2 \) is then equal to \( a \)
or \( b \) only. There are thus 106 word lists. The data- 
base being very large (totaling 23 hours of speech,
\( \mu \)-law coded), we defined the training set, the cross-
validation set and the test set as follows:

- a “reduced” training set totaling approximately
  5 hours of speech:
  all \( sa, sh, sm, sq \), and \( st \) word lists, i.e., 21
  word lists.

- the cross-validation set (use to adapt the MLP
  learning rate):
  all \( so \) and \( sy \) word lists (8 word lists).

- the test set:
  all \( sd \) and \( sr \) word lists, i.e., 8 word lists. Since
  the lexicon is different in each of these 8 word
  lists, we then have the choice to recognize the
  8 word lists as a whole (yielding a lexicon of
  600 words) or to recognize each word list in-
  dependently with a lexicon of about 75 words.

In the second case, the recognition rate will be
the (unweighted) average over the 8 recogni-
tion rates.

To generate the phonetic transcriptions of the
training words as well as of the test words, we have
used the 110,000-word CMU 0.4 dictionary defin-
ing 46 phonemes (a subset of the TIMIT phone set).
Some of the Phonebook words were not present in
CMU 0.4 and were transcribed manually.

4. EXPERIMENTS

Experiments have been done to compare context in-
dependent (CI) models with context independent
models combined with context dependent (CD) mod-
els (CI&CD). A minimum duration of half the aver-
age duration of each phoneme was used to define the
CI model topologies. The ANNs were feed-forward
multilayer perceptrons. The number of parameters
was kept to 166,000 for all of the compared systems.

We have been using lpc-cepstral features with
cepstral mean subtraction (CMS). These features
have been chosen for their robustness against channe-
l and speaker characteristics. These parameters
were computed every 10 ms on 30 ms analysis win-
dows. The order of the lpc analysis was set to 10.

The complete feature set for our hybrid
HMM/ANN systems was based on a 26 dimensional
vector composed of the cepstral parameters (log-
RASTA-PLP or lpc-cepstral parameters with cep-
stral mean subtraction), the Δcepstral parameters,
the Δenergy and the ΔΔenergy. Nine frames of
contextual information was used at the input of the
ANNs, leading to 234 inputs.

Three tying configurations were compared. For
the CI&CD(I) case, generalized transitions were based
on the phonetic classes (9 classes + silence) of the
left and right phonemes of the transition. Hence,
the neural network estimating context independent
probabilities as well as context dependent probabili-
ties had 46 (CI phones) + \( 10 \times 10 \) (CD transitions) =
146 outputs. For the CI&CD(II) case, generalized
transitions were based on the place of articulation
of the left and right phonemes of the transition (8
classes + silence). The neural network had 127 out-
puts. Finally, for the CI&CD(III) case, we used the
automatic data driven clustering approach to obtain
a set of 81 transition classes (neural network with
127 outputs).
5. RESULTS

<table>
<thead>
<tr>
<th>Model type</th>
<th>75 words</th>
<th>600 words</th>
<th>8000 words</th>
</tr>
</thead>
<tbody>
<tr>
<td>CI</td>
<td>1.5%</td>
<td>5.3%</td>
<td>17.3%</td>
</tr>
<tr>
<td>CI&amp;CD(I)</td>
<td>1.1%</td>
<td>3.9%</td>
<td>13.8%</td>
</tr>
<tr>
<td>CI&amp;CD(II)</td>
<td>1.1%</td>
<td>3.3%</td>
<td>12.0%</td>
</tr>
<tr>
<td>CI&amp;CD(III)</td>
<td>1.1%</td>
<td>4.2%</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 3: Error rates on isolated word recognition (75, 600 and 8000 lexicon words) with hybrid HMM/ANN systems and CMS features. Comparison between context independent and combination of context independent and context dependent models. The 3 kinds of phone to phone transition clustering are presented here.

As shown in Table 3, using CD&CI models yielded up to 38% reduction of the error rate with only a slight increase of the computational load due to the modified word topologies. All the experiments were done with the STRUT (Speech Training and Recognition Unified Toolkit) software [1]. So far the best performance were obtained with the phonetic classes based on the place of articulation. However, we intend to further investigate the data-driven clustering (CD(III)) since this approach does not require any a priori knowledge about the phoneme characteristics.

6. CONCLUSIONS

In this paper, we have presented a particular hybrid HMM/ANN configuration in which we combine context independent phone models and context dependent generalized diphone models. Three kinds of phone to phone transition clustering have been compared in the framework of a training independent task. Experiments have demonstrated significant improvements on both small and medium size vocabulary. This first attempt to use CD phone models with transition models in a hybrid architecture looks promising (up to 38% improvement). Future work will consist in applying this approach to other tasks (continuous speech recognition, keyword spotting, ...). On another side, different clustering methods could be investigated in order to define the phonetic classes: neural networks, tree based clustering, ...

7. ACKNOWLEDGMENTS

We thank the European Community for their support in this work (SPRACH Long Term Research Project 2007).

8. REFERENCES

http://tcts.fpms.ac.be/speech/strut.html


