A NEW METHOD TO ACHIEVE FAST ACOUSTIC MATCHING FOR SPEECH RECOGNITION

Clark Z. Lee and Douglas O'Shaughnessy

INRS-Télécommunications
16 Place du Commerce, Verdun, Quebec, Canada H3E 1H6

ABSTRACT

For large vocabulary continuous speech recognition based on hidden Markov models, we often face the issue of trade-off between the accuracy and the speed. A new method is proposed in this article such that complex models are used to retain a high accuracy whereas the speed is achieved by using the similarities in acoustic matches. These similarities are based on the assumption that we refer as a look-phone-context property. By using the look-phone-context property, the number of acoustic matches can be substantially reduced in the course of scoring all possible phonetic transcriptions of recognition hypotheses. Experiments on the speaker-independent *Wall Street Journal* task show that a fast-response system can be reached without compromising the accuracy.

1. INTRODUCTION

For large vocabulary continuous speech recognition, we have adopted a two-pass search strategy in which inexpensive models are used in the first pass to produce the word graph [1, 5] and more powerful language models and acoustic models are used in the second pass to rescore the word graph. In the article [4], we developed a clustering method to increase the accuracy, which considered both left and right phone contexts as well as word-level information, the beginning and ending of words and function words. Since the number of acoustic distributions of the second pass after clustering is about 8,000 in the 30,000-word system, which is nearly tripled compared with that in our previous 5,000-word real-time system [6], a new approach in the second pass is necessary to achieve a fast-response system.

The objective of the second pass is to find the highest scoring recognition hypothesis by searching the word graph produced by the first pass. We allow for multiple segmentation hypotheses [6] in order to score partial transcriptions exactly, whereas in the first pass, each partial transcription that is hypothesized has a unique segmentation associated with it. We use a depth first search algorithm to conduct the search. As far as the acoustic matching is concerned, the principal operation is to propagate an array of forward scores for each phone with its contexts. A naive implementation would require an acoustic match for each phone of each phonetic transcription to be scored in the course of generating recognition hypotheses. The idea of using the similarities of acoustic matches is based on the following observations. We cannot expect that we will obtain the same acoustic matches by propagating any different arrays of forward scores. However, it is reasonable to assume that if two arrays of forward scores are centered at the same time and are propagated through the same acoustic model, they will give rise to approximately identical acoustic matches provided that for some suitable integers l and r, their two phonetic strings of recognition hypotheses agree on l look-left phones and r look-right phones. The above property is referred as look-phone-context property.

Assuming the look-phone-context property holds, we do not have to calculate an acoustic match every time a phone is to be scored. This is achieved by constructing a transcription graph which encodes phonetic transcriptions and the information of multiple segmentations. A path in the transcription graph is created only if it does not exist and if its corresponding acoustic matches are requested to generate a new recognition hypothesis. The construction of the transcription graph is using a depth first search and is guided by the language model. Notice that there is a many-to-one correspondence between paths through the recognition hypotheses and paths through the transcription graph. Accordingly, the acoustic matches encoded in the transcription graph can be re-used in the course of generating recognition hypotheses.

2. THE GENERATION OF RECOGNITION HYPOTHESES

Let us take a look at the details of generating recognition hypotheses in the second pass. Since the search space in the second pass is the word graph produced in the first pass, all we need to do is to rescore the word graph. The specification of the word graph is the following. A node is labeled by a triple $(\bar{t}, \mathbf{D}, \bar{\sigma})$ where $\bar{t}$ is a first pass time, $\mathbf{D}$ is a look-ahead phone string and $\bar{\sigma}$ is a coarse language model state [5]. A branch joining two nodes is associated with a complete lexical theory. A lexical theory is a partial transcription (or a node in the lexical tree) together with a phone segmentation where it is said to be complete when it reaches a leaf node of the lexical tree. The word graph is constructed by using the monotone graph search algorithm [2, 5].

To carry out the second pass search, we define a partial recognition hypothesis to be a partial path in the word graph together with the information needed to support scoring with the fine acoustic phonetic models and the fine language models. The partial recognition hypothesis η is a quadruple $(b, \mathbf{E}, \{\alpha_t\}, \{\Lambda_\sigma\})$ where

1. b is a branch in the word graph.

2. $\mathbf{E}$ is a look-behind phone string.

3. $\{\alpha_t\}$ is an array of forward scores centered on $\bar{t}$ whose width is controlled by the uncertainty Δ.

4. $\{\Lambda_\sigma\}$ is an array of language model scores indexed by fine language model states σ.

Suppose b is a branch in the word graph originating in a node $(\bar{t}, \mathbf{D}, \bar{\sigma})$ and terminating in another node $(\bar{t}', \mathbf{D}', \bar{\sigma}')$. Let w be the word label and $\mathbf{F}$ the phonetic transcription of w in the complete lexical theory corresponding to b. We use the branch b to generate a new partial recognition hypothesis $\eta' = (b', \mathbf{E}', \{\alpha'_{t'}\}, \{\Lambda'_{\sigma'}\})$ as follows:

1. $\mathbf{E}'$ is the look-behind phone string defined by the condition that the tail end of $\mathbf{DF}$ is $\mathbf{E}'\mathbf{D}'$.

2. $\{\alpha'_{t'}\}$ is the array of forward scores centered on $\bar{t}'$ obtained by propagating $\{\alpha_t\}$ through the phone string $\mathbf{DF}\backslash\mathbf{D}'$ (that is, $\mathbf{DF}$ with $\mathbf{D}'$ removed).

3. For each language model state σ' such that $P(w, \sigma'|\sigma) > 0$ for some state σ in η,
$$\Lambda'_{\sigma'} = \max_{\sigma} \Lambda_\sigma P(w, \sigma'|\sigma).$$

3. THE LOOK-PHONE-CONTEXT PROPERTY

The role of the acoustic matching in generating a new partial recognition hypothesis is to propagate the array of forward scores through the phone string of the corresponding branch in the word graph. This can be done by consecutively propagating through one phone at a time. Suppose we have an input array $\{\alpha_t\}$ of forward scores centered on $\bar{t}$ with a look-left phone string $\mathbf{L}$ and a look-right phone string $\mathbf{R}$. After propagating through a phone, we get an output array $\{\alpha'_{t'}\}$ of forward scores centered on $\bar{t}'$ with a look-left phone string $\mathbf{L}'$ and a look-right phone string $\mathbf{R}'$, where $\bar{t}, \bar{t}'$ are the first pass phone segmentation. We may calculate an array of phone durations $\{d_{t'}\}$ and an array of phone scores $\{s_{t'}\}$ centered on $\bar{t}'$. We say the look-phone-context property holds if the durations $\{d_{t'}\}$ and phone scores $\{s_{t'}\}$ are the same for two different acoustic matches with two different input arrays of forward scores, but with the same first pass segmentation $\bar{t}, \bar{t}'$ and the same look-left and look-right phone strings $\mathbf{L}, \mathbf{R}, \mathbf{L}', \mathbf{R}'$. Note that in order to ensure the phone contexts are always available during the procedure, the number of phones in the look-behind phone string $\mathbf{E}$ in the partial recognition hypothesis must be equal or larger than the number of phones in $\mathbf{L}$, and the number of phones in the look-ahead phone string

$\mathbf{D}$ in the partial recognition hypothesis must be equal or larger than the number of phones in $\mathbf{R}$.

4. THE TRANSCRIPTION GRAPH

The transcription graph is constructed based on the assumption that the look-phone-context property approximately holds for some suitable lengths of look-left and look-right phone strings. The transcription graph is defined such that there is a many-to-one correspondence between paths through the word graph and paths through the transcription graph. We specify a node in the transcription graph by means of a triple $(\bar{t}, \mathbf{L}, \mathbf{R})$, where $\bar{t}$ is the first pass segmentation and $\mathbf{L}, \mathbf{R}$ are look-left and look-right phone strings. The information of multiple segmentations is encoded on branches. A branch from a node $(\bar{t}, \mathbf{L}, \mathbf{R})$ to another node $(\bar{t}', \mathbf{L}', \mathbf{R}')$ is associated with an array of durations $\{d_{t'}\}$ and array of phone scores $\{s_{t'}\}$.

The construction of the transcription graph is guided by the language model, namely nodes and branches are only created if they do not exist and if such acoustic matches are requested in order to generate a new partial recognition hypothesis. We use a depth first search to rescore the word graph. Before the search starts, we order the branches of the word graph according to their combining forward and backward scores since that allows us to get the optimal path earlier [6]. The basic operation of the acoustic matching is to add a phone to an input array $\{\alpha_t\}$ and to generate an output array $\{\alpha'_{t'}\}$. To score a first pass phone segment $(\bar{t}, \mathbf{L}, \mathbf{R}), (\bar{t}', \mathbf{L}', \mathbf{R}')$, we do the following in the transcription graph:

1. Create nodes $(\bar{t}, \mathbf{L}, \mathbf{R})$, $(\bar{t}', \mathbf{L}', \mathbf{R}')$ if they do not exist in the transcription graph.

2. If the corresponding branch does not exist in the transcription graph, calculate $\{\alpha'_{t'}\}$ by propagating $\{\alpha_t\}$ and keep an array of back pointers on the entry times $\{q_{t'}\}$, then calculate an array of durations $\{d_{t'}\}$ and array of phone scores $\{s_{t'}\}$ and create the branch,

$$\begin{aligned} d_{t'} &= t' - q_{t'} + 1, \\ s_{t'} &= \alpha'_{t'} - \alpha_{q_{t'}-1}. \end{aligned}$$

3. If the corresponding branch exists, calculate the output array $\{\alpha'_{t'}\}$,

$$\begin{aligned} \tau &= t' - d_{t'}, \\ \alpha'_{t'} &= \alpha_\tau + s_{t'}. \end{aligned}$$

In the course of building the transcription graph, we also impose envelope pruning. Nodes and branches are created only when they survive during the pruning. There are two pruning envelopes, one for word boundaries and another for phone boundaries, which keep track of the best forward scores for each frame. Separate thresholds are used against the envelopes for pruning phones and words since the forward score used to calculate the word envelope includes the language model score. Pruning is also carried out after finishing searching a block, in order to prevent unnecessary recognition hypotheses from passing to the next block.

Since the nodes and branches in the transcription graph are created on demand, the transcription graph only contains the paths required for rescoring the word graph. On the other hand, since the transcription graph contains no language model information, it allows the second pass search to be nearly independent of the size of the language model.

5. EXPERIMENTAL RESULTS

To test the system, we used the *Wall Street Journal* (WSJ) speaker-independent corpus with the SI284 training set to train gender-dependent acoustic models. Acoustic features were calculated every 10 ms from the 16 kHz sampled data after DC-component removal. The feature vector consisted of 15 cepstral coefficients, 15 delta and 15 delta delta coefficients, where a simple mean normalization was imposed on a fixed window basis. Clustering by using decision trees was applied for both first and second pass models. VQ models with 3 codebooks and with only right contexts were used in the first pass, where each codebook consisted of one covariance matrix, 256 means and a set of distributions. For the second pass, the acoustic models had 2 codebooks, each of which had one grand covariance matrix and up to 16 means per distribution. The clustering was on triphone contexts as well as high-level knowledge sources, the beginning and ending of words and function words. After clustering there were about 8,000 output distributions for each gender. The details of the clustering method can be found in [4].

The language models were derived from the statistics of *North American Business* texts provided by CMU. The vocabulary was chosen such that the most frequent 30,000 words according to the unigram statistics intersected with the COMLEX dictionary, which resulted in 29,533 words. With this vocabulary, we obtained about 5.5 million bigrams and 6.3 million trigrams, where count-1 statistics were excluded for bigrams and count-3 and below were excluded for trigrams. We used bigrams in the first pass and trigrams in the second pass.

For the first pass search, time quantization was imposed to reduce the number of nodes in the word graph. For lexical access, we used a language model heuristic, depth first search and local envelope pruning, which were shown to be efficient for large vocabulary applications [3].

We have performed an open-vocabulary test on the WSJ evaluation set (Nov92-20k-si-nvp) of the speaker-independent continuous speech recognition task. The out-of-vocabulary (OOV) rate was 1.2% with the 29,533-word vocabulary. We have carried out a contrast experiment to validate the look-phone-context property and to see the difference between the baseline system and the system with the transcription graph. We have not observed any degradation of recognition accuracy in the 30k-word WSJ task if two-phone look-left and two-phone look-right are used in building the transcription graph. On the other hand, we increase the speed by about three times for the second pass rescoring. The whole system currently runs in faster than real-time on a Sun workstation (with clock speed 170 MHz) with a word error rate 11.36%.

6. CONCLUSION

We have presented a new method of using the look-phone-context property to achieve fast acoustic matching in the second pass of our speech recognition system. The sharing of acoustic matches is accomplished by constructing a transcription graph which contains only the paths of phonetic transcriptions required for rescoring the word graph. Experimental results suggest that this approximate method is robust and effective for large vocabulary continuous speech recognition.

ACKNOWLEDGMENTS

This work was supported by the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

[1] H. Ney, S. Ortmanns and I. Lindam, “Extensions to the word graph method for large vocabulary continuous speech recognition,” *Proceedings ICASSP 97*, vol. 3, pp. 1791–1794, April 1997.

[2] N. Nilsson, *Principles of Artificial Intelligence*, Tioga Publishing Company, 1980.

[3] C.Z. Lee and D. O’Shaughnessy, “Techniques to achieve fast lexical access,” *IEEE ASRU Workshop 97*, Santa Barbara, California, December 1997.

[4] C.Z. Lee and D. O’Shaughnessy, “Clustering beyond phoneme contexts for speech recognition,” *Proceedings Eurospeech 97*, vol. 1, pp. 19–22, September 1997.

[5] Z. Li, G. Boulianne, P. Labute, M. Barszcz, H. Garudadri and P. Kenny, “Bi-directional graph search strategies for speech recognition,” *Computer Speech and Language*, vol. 10, pp. 295–321, 1996.

[6] Z. Li, M. Heon and D. O’Shaughnessy, “New developments in the INRS continuous speech recognition system,” *Proceedings ICSLP 96*, vol. 1, pp. 2–5, October 1996.