IMPROVING SPEECH RECOGNITION BY EXPLICIT MODELING OF PHONE DELETIONS

Tom Ko, Brian Mak

Department of Computer Science and Engineering
The Hong Kong University of Science and Technology
Clear Water Bay, Hong Kong
{tomko, mak}@cse.ust.hk

ABSTRACT

In a paper published by Greenberg in 1998, it was said that in conversational speech, phone deletion rate may go as high as 12% whereas syllable deletion rate is about 1%. The finding prompted a new research direction of syllable modeling for speech recognition. To date, the syllable approach has not yet fulfilled its promise. On the other hand, there were few attempts to model phone deletions explicitly in current ASR systems. In this paper, fragmented word models were derived from well-trained cross-word triphone models, and phone deletion was implemented by skip arcs for words consisting of at least four phonemes. An evaluation on CSR-II WSJ1 Hub2 5K task shows that even with this limited implementation of phone deletions in read speech, we obtained a word error rate reduction of 6.73%.

Index Terms— Phone deletions, acoustic modeling, fragmented word model, skip arc, syllable

1. INTRODUCTION

In his seminal papers [1, 2], Greenberg presented a syllabic-centric perspective for understanding pronunciation variation. One of his major findings from a systematic analysis of manually transcribed conversations from the Switchboard [3] corpus is that syllable is a more stable linguistic unit for pronunciation modeling than phoneme: (a) in Switchboard, phone deletion rate is about 12% whereas syllable deletion rate is about 1%; (b) syllable onsets are well preserved; syllable nuclei may change; syllable coda are frequently disposed.

The findings prompted a new research direction in the automatic speech recognition (ASR) community to investigate the modeling of syllables as the acoustic units for ASR [4, 5, 6, 7, 8, 9, 10], or to incorporate syllable information to improve speech recognition [11, 12]. To date, the gain from syllable modeling on ASR is modest (if there is any at all) when compared with a standard cross-word triphone-based system.

This paper is not another attempt of syllable modeling. Instead, we would like to investigate the effectiveness of incorporating phone deletions explicitly in a conventional cross-word triphone-based system. In the past research of syllable-based ASR, the research efforts focused mainly on (a) how to determine the mixing of phone units with syllable units [7, 13]? (b) How to model context dependency in syllable modeling without an explosion of units [9, 10]? (c) How to solve the data sparsity problem due to the explosion of syllable units, especially when context-dependent syllables were used [7, 9, 10]?

However, phone substitutions and phone deletions were not explicitly modeled except for a few failed attempts. For instance,

- In [13], skip arcs were added to some syllable states, but the purpose is not to model phone deletions but to downplay states that were not reliably trained. However, state skipping resulted in performance degradation.
- In [8], multi-path syllable models were investigated to model pronunciation variations but again resulted in poorer ASR performance.

On the other hand, Jurafsky designed an interesting experiment (again on Switchboard) to investigate what kinds of pronunciation variations was hard for triphone modeling [14] using additional training data with canonical lexicon. It turned out that the current method of triphones training could model phone substitution and vowel reduction quite well, but had problem with modeling syllable deletions. Inspired by both Greenberg's and Jurafsky's findings, in this paper, we investigate explicit modeling of phone deletions[1] in (whole) word models that are bootstrapped from cross-word triphone models. For the CSR-II WSJ1 Hub2 5K recognition task, even when we limit our investigation to words consisting of at least 4 phonemes, we are surprised that phone deletion modeling reduces word error rate by an absolute 0.57% or a relative of 6.73%.

This paper is organized as follows. Explicit modeling of phone deletions is described in the next section, which is followed by the experimental evaluation in Section 3 and Conclusions in Section 4.

2. EXPLICIT MODELING OF PHONE DELETIONS

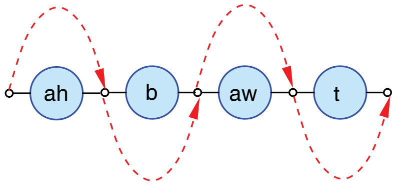

Fig. 1. An example of adding skip arcs to allow phone deletions.

The idea of allowing phone deletion by skip arcs as shown in Fig. 1 is simple. In practice, since we use existing trainer and decoder (and in our case, HTK's), we have to choose a linguistic unit larger than a phoneme for its implementation. From the research results of syllable modeling, we learn that in acoustic modeling,

This work was supported by the Research Grants Council of the Hong Kong SAR under the grant numbers HKUST617008 and HKUST617507.

[1] Phone deletions may be considered as more general than just syllable deletions. In some cases, a single phone or a sequence of phone deletions is equivalent to a syllable deletion.

978-1-4244-4296-6/10/$25.00 ©2010 IEEE

Table 1. Examples of context-dependent fragmented word model (where '?' represents any phone in the actual context).

Word	Phonemic Transcription	Modified Transcription	Context-dependent Fragmented Word Model
about	ah b aw t	ah b^aw t	?-ah+b^aw ah-b^aw+t b^aw-t+?
consider	k ah n s ih d er	k ah n^s^ih d er	?-k+ah k-ah+n^s^ih ah-n^s^ih+d n^s^ih-d+er d-er+?

- context dependency modeling is important. However, if the acoustic unit is bigger than a phoneme, the number of context-dependent models for such unit (like syllable) may be astronomical.
- a tradeoff must be made between the number of acoustic units and available training data.

Inspired by the work in [7, 10], we propose *context-dependent fragmented (whole) word models* (CD-FWM) to implement phone deletions, which are constructed from well-trained cross-word triphones.

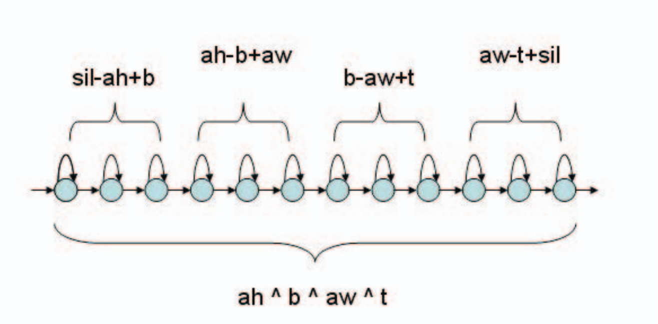

Fig. 2. An example of the construction of a context-independent word model from word-internal triphones for the word "about".

2.1. Context-dependent Fragmented Word Models (CD-FWM)

A context-independent (CI) word model may be easily constructed from word-internal triphones as shown in Fig. 2 for the word "about". The recognition performance of CI word models constructed in this way should be the same as that of the original word-internal triphones. However, modeling contextual word models is not easy, and a naive approach of "tri-word modeling" is infeasible even for a modest task with a few hundred words in its vocabulary.

Following the approach of fragmented context-dependent syllable models in [10], we propose the *context-dependent fragmented word models (CD-FWM)* and split a word into three or more segments so that the center segment is not influenced by cross-word contexts. This will greatly reduce the number of of possible context-dependent units. To further contain the number of CD acoustic units in CD-FWM, the number of segments depends on the length L of a word, which is defined as the number of phonemes in its canonical pronunciation, as follows:

- $L \leq 3$: the word is represented by the original cross-word triphones instead of a word model, and no phone deletions are allowed.
- $L = 4$ or 5: the word is split into 3 segments with the first and the last segment consisting of a single phone. Table 1 gives an example of a 3-segment CD-FWM for the word "about".
- $L \geq 6$: the word is split into 5 segments with the first two and the last two segments consisting of a single phone. Table 1 gives an example of a 5-segment CD-FWM for the word "consider".

Table 2. Coverage of words of various lengths in the WSJ lexicon and corpora.

Word Length	WSJ 5K Lexicon	Hub2 Eval Word Tokens
$L \geq 6$	2,672 (54%)	942 (26%)
$L \geq 4$	4,314 (86%)	1,817 (50%)
$L \geq 1$	4,989 (100%)	3,647 (100%)

The coverage of words of various lengths is shown in Table 2.

Thus, in a CD-FWM, there are actually both CD phone units and CD subword units (SWU). In a 3-segment CD-FWM, both the first and the last segments are affected by cross-word contexts, and they are not the conventional triphones: the right context of the first segment, and the left context of the last segment is the center subword segment. (We call them additional CD phones as they are not the conventional triphones.) On the other hand, the first and the last segment for a 5-segment CD-FWM are just original cross-word triphones; the middle three segments are similar to a 3-segment CD-FWM. The important point is that for words with $L \geq 4$, the center SWU is almost unique for each word. As a consequence, the number of acoustic units only increases by $O(nV)$, where n is the number of phonemes and V is the size of the vocabulary, instead of $O(V^3)$ if "tri-words" are used.

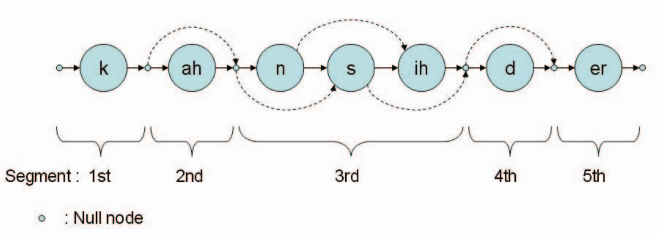

Fig. 3. An example of adding skip arcs to allow phone deletions in the actual implementation of context-dependent fragmented word models (CD-FWM) for the word "consider".

2.2. Practical Implementation of CD-FWM

In the practical implementation of CD-FWM by HTK, right now we cannot skip two successive phones within an SWU. The reason is that an SWU is represented by an HMM unit, and there are no null nodes inside an HMM unit in HTK. To skip the first phone of an SWU, a skip arc is constructed from the previous null node to the first state of its second phone. To skip a middle phone in an SWU, a skip arc is added to jump from the last state of its previous phone to the first state of its following phone. To skip the last phone of an SWU, a skip arc is added to jump from the last state of its previous phone to the following null node. Fig. 3 shows an example with the word "consider".

We further did not allow skipping the first and the last phone in a CD-FWM. Skipping them may increase the number of context-dependent units drastically.

Table 3. Recognition performance on Nov'93 Hub2 5K evaluation task. All models have 5864 tied states. (SWU = Sub-Word Units)

Model	#CD Phones	#CD SWUs	Word Accuracy
cross-word triphones	62,402	0	91.53%
CD-FWM for $L \geq 6$	79,767	9,117	91.55%
CD-FWM for $L \geq 6$ + phone deletion	79,767	9,117	92.10%
CD-FWM for $L \geq 4$	380,341	13,907	91.58%
CD-FWM for $L \geq 4$ + phone deletion	380,341	13,907	92.05%

3. EXPERIMENTAL EVALUATION ON CSR-II WSJ1 HUB2 5K RECOGNITION TASK

The effectiveness of modeling phone deletions in LVCSR using the proposed context-dependent fragmented word models (CD-FWM) was evaluated on the CSR-II WSJ1 Hub2 5K recognition task.

Table 4. Information of various WSJ data sets.

Data Set	#Speakers	#Utterances	Vocab Size
train (si_tr_s)	302	46,995	13,725
dev1 (si_et_05)	8	330	1,270
dev2 (si_dt_05)	10	496	1,842
eval (si_et_h2)	10	205	998

3.1. Speech Corpora

Conventional speaker-independent cross-word triphone models and the proposed CD-FWMs were trained on the standard SI-284 WSJ training data plus additional WSJ adaptation data and short-term training data in the WSJ0 and WSJ1 corpora. It consists of 8,720 WSJ0 utterances from 101 WSJ0 speakers and 38,275 WSJ1 utterances from 201 WSJ1 speakers. Thus, there is a total of about 44 hours of read speech in 46,995 training utterances from 302 speakers.

The standard Nov'93 5K non-verbalized Hub2 test set si_et_h2 was used for evaluation using the standard 5K-vocabulary bigram that came along with the WSJ corpus. The optimal decoding parameters were only tuned on the baseline cross-word triphones system by grid search using the WSJ1 5K development set si_dt_05. They were then simply adopted for all other systems under investigation. Notice that utterances containing OOV words were removed from both the development and evaluation test sets. A summary of these data sets is shown in Table 4.

3.2. Training of Cross-word Triphone Models

The SI baseline model consists of 62,402 cross-word triphones based on 39 base phonemes. Each triphone model is a strictly left-to-right 3-state continuous-density hidden Markov model (CDHMM), with a Gaussian mixture density of at most 16 components per state, and there are totally 5,864 tied states. (The model complexity was tuned using another development set si_et_05.) In addition, there are a 1-state short pause model and a 3-state silence model.

The traditional 39-dimensional MFCC vectors were extracted at every 10ms over a window of 25ms.

The models were first trained on the training utterances with no endpoint detection. This first set of models were then used to endpoint all training data, and they were then re-trained with the endpointed data.

3.3. Training of Context-dependent Fragmented Word Models (CD-FWM)

CD-FWM were derived from the baseline cross-word triphones as follows:

STEP 1 : The canonical pronunciation of each word in the dictionary was modified: the original phonemic representation was replaced by the corresponding FWM segments. Note that the number of segments in the FWM of a word depends on its length as described in Section 2.1. The number of cross-word triphones, additional CD phones, and new CD subword units (SWU) in the CD-FWMs for different settings are shown in Table 3. Notice the 5-fold increase in the number of CD phones when CD-FWMs were constructed for words consisting of at least four phonemes vs. six phonemes.

STEP 2 : The required models in the CD-FWM system: cross-word triphones, additional CD phones, and CD SWUs were then constructed from the cross-word triphones in the baseline system. At this point, the two systems are basically the same — with the same set of tied states (and, of course, the same state-tying structure) — and have the same recognition performance.

STEP 3 : Skip arcs were added to the additional CD phones and CD SWUs to allow deletion of any phones in a word except the first one and the last one.

STEP 4 : The new CD-FWMs with skip arcs were re-trained for four EM iterations.

As a sanity check for the efficacy of phone deletions, we also re-trained the models constructed from STEP 2 without adding the phone deletion skip arcs for four EM iterations in another experiment. Notice that although the underlying tied states in CD-FWMs are the same as those in the cross-word triphones that derive them, due to the SWUs (which are represented by the center segments in the FWMs), after re-training the acoustic models that involve those center segments (e.g., ?-ah+b^aw in Table 1) will have their own state transitions different from those in the original triphones, and they are almost word-dependent (because only a few words will share these units which have a context spanning over more than three phonemes). The state distributions might also be different after re-training.

3.4. Results and Discussion

The recognition performance of the cross-word triphone baseline and the various CD-FWM systems are shown in Table 3. It can be seen that without the addition of phone deletion skip arcs, re-trained CD-FWMs give almost no recognition improvement over the baseline triphone system[2]. Although the new CD phones and CD SWUs

[2] We had empirically verified, as expected, that CD-FWMs gave the same recognition performance as the baseline triphones which derived them if they were not re-trained.

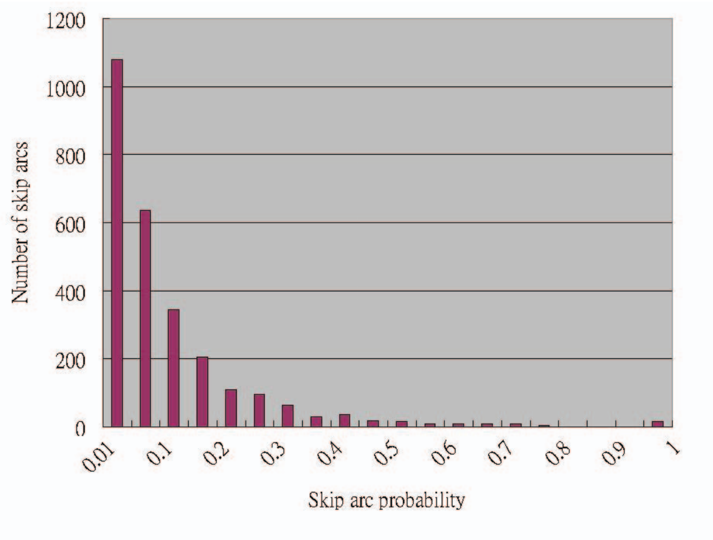

Fig. 4. Distribution of phone deletion probabilities for the CD-FWM system with $L \geq 6$. Those with a probability less than 0.01 are removed from this plot.

in CD-FWMs may model some word-specific information through the re-estimated state transitions in those models, since state transitions are much less important than the state distributions in an HMM, the improvement is expected to be small, if any.

The biggest gain comes from the addition of skip arcs to allow phone deletions; it is 0.57% absolute. We are disappointed by the result from $L \geq 4$ which is not better than that from $L \geq 6$.

3.5. Analysis of the Skip Arc Probabilities

We looked at the estimated probabilities of the phone deletion skip arcs for the CD-FWM system that implemented phone deletions for words with six or more phonemes. A distribution of their probabilities is plotted in Fig. 4. Out of the total 52,507 phone deletion skip arcs, 49,814 (about 95%) of them have a probability less than 0.01 (which are not included in the plot of Fig. 4). Thus, only about 5% of the phones (in the added CD phones and CD SWUs) benefit from possible deletions; yet, the recognition improvement is relatively substantial.

We will examine these 5% phone deletions in greater detail in our future work.

4. CONCLUSIONS

In this paper, we hypothesize that phone deletions, though are more common in spontaneous speech, may also occur in read speech, and investigated their effectiveness on the WSJ task. They were implemented by the concept of *context-dependent fragmented word models* (CD-FWM) which were composed from well-trained cross-word triphones. As a consequence, their recognition performance should not be worse than that of cross-word triphones. To contain the expansion of subsequent context-dependent phone units and subword units, we limited CD-FWMs for words consisting of four or more phonemes. For the remaining words, they were modeled by triphones. On the Hub2 evaluation set, the word recognition accuracy improved from the baseline 91.53% (given by cross-word triphones) to 92.10%.

The construction of CD-FWMs can model some word-specific information. Right now, although both state transitions and state distributions in the CD-FWMs were re-estimated, only the re-estimated state transitions may capture word-specific information because the state distributions were still shared with the global triphones since the original tied states were kept. We would like to investigate how to untie some of these tied states to better model word-specific information. Moreover, in the current implementation, we could not delete the last phoneme in a word; this is obviously not correct. Future work will deal with all these limitations using spontaneous speech.

5. REFERENCES

[1] S. Greenberg, "Understanding speech understanding towards a unified theory of speech perception," in *Proceedings of the ESCA Tutorial and Advanced Research Workshop on the Auditory Basis of Speech Perception*, W.A. Ainsworth and S. Greenberg, Eds. 1996, pp. 1–8, Keele University, UK.

[2] S. Greenberg, "Speaking in shorthand — a syllable-centric perspective for understanding pronunciation variation," in *Proceedings of the ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition*, 1998.

[3] J. Godfrey, E. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *Proc. of ICASSP*, Mar. 1992, pp. 517–520.

[4] A. Hauenstein, "Using syllable in a hybrid HMM-ANN recognition system," in *Proc. of Eurospeech*, 1997, pp. 1203–1206.

[5] R. J. Jones, S. Downey, and J. S. Mason, "Continuous speech recognition using syllables," in *Proc. of Eurospeech*, 1997, vol. 3, pp. 1171–1174.

[6] A. Ganapathiraju, J. Hamaker, J. Picone, M. ordowski, and G. Doddington, "Syllable-based large vocabulary continuous speech recognition," *IEEE Trans. on SAP*, vol. 9, no. 4, pp. 358–366, May 2001.

[7] A. Sethy and S. Narayanana, "Split-lexicon based hierarchical recognition of speech using syllable and word level acoustic units," in *Proc. of ICASSP*, 2003, vol. I, pp. 772–775.

[8] A. H*ä*m*ä*l*ä*inen, L. Bosch, and L. Boves, "Construction and analysis of multiple paths in syllable models," in *Proc. of Interspeech*, 2007, pp. 882–885.

[9] Hao Wu and Xihong Wu, "Context dependent syllable acoustic model for continuous Chinese speech recognition," in *Proc. of Interspeech*, 2007, pp. 1713–1716.

[10] K. thambiratnam and F. Seide, "Fragmented context-dependent syllable acoustic models," in *Proc. of Interspeech*, 2008, pp. 2418–2421.

[11] Su-Lin Wu, M. Shire, S. Greenberg, and N. Morgan, "Integrating syllable boundary information into speech recognition," in *Proc. of ICASSP*, 1997, vol. 2, pp. 987–990.

[12] Su-Lin Wu, E. Kingsbury, N. Morgan, and S. Greenberg, "Incorporating information from syllable-length time scales into automatic speech recognition," in *Proc. of ICASSP*, 1998, vol. 2, pp. 721–724.

[13] A. Sethy, B. Ramabhadran, and S. Narayanana, "Improvements in English ASR for the MALACH project using syllable-centric models," in *IEEE ASRU*, 2003.

[14] D. Jurafsky, W. Ward, J. P. Zhang, K. Herold, X. Y. Yu, and S. Zhang, "What kind of pronunciation variation is hard for triphones to model?," in *Proc. of ICASSP*, 2001.