Problems of Modeling Phone Deletion in Conversational Speech for Speech Recognition

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

***Abstract*—Recently we proposed a novel method to explicitly model the phone deletion phenomenon in speech, and introduced the *context-dependent fragmented word model* (CD-FWM). An evaluation on the WSJ1 Hub2 5K task shows that even in read speech, CD-FWM could reduce word error rate (WER) by a relative 10.3%. Since it is generally expected that the phone deletion phenomenon is more pronounced in conversational and spontaneous speech than in read speech, we extend our investigation of modeling phone deletion in conversation using CD-FWM on the SVitchboard 500-word task in this paper. To our surprise, much smaller recognition gain is obtained. Through a series of analyses, we present some plausible explanations for why phone deletion modeling is more successful in read speech than in conversational speech, and suggest future directions in improving CD-FWM for recognizing conversational speech.**

I. Introduction

Phone deletion is a common phenomenon in human speech. In [1], [2], Greenberg performed a systematic analysis of manually transcribed conversations from the Switchboard corpus [3] and found that phone deletion rate is about 12% in the corpus. It is plausible that speech recognition performance can be greatly improved if the phone deletion phenomenon is effectively modeled. However, the problem is not much addressed in today's acoustic models.

There were some failed attempts in modeling phone variations that had also indirectly modeled phone deletion to a certain extent. For instance,

- In [4], 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 [5], multi-path syllable models were investigated to model pronunciation variations but again resulted in poorer ASR performance.

On the other hand, Jurafsky [6] empirically found that, on Switchboard, the current method of triphones training could model phone substitution and vowel reduction quite well, but had problem with modeling syllable deletion.

Recently, we proposed the *context-dependent fragmented word model* (CD-FWD) to model phone deletion[1] explicitly [7]. CD-FWMs are whole word models that are bootstrapped from tied-state cross-word triphone models. We first tested CD-FWM on read speech on the CSR-II WSJ1 Hub2 5K recognition task, and were surprised that even in read speech, CD-FWM could reduce word error rate (WER) by an absolute 0.87% or a relative of 10.3% as shown in Table I[2]. In this paper, we extended our investigation of using CD-FWM to model phone deletion in the SVitchboard conversational speech. Since the phone deletion phenomenon is more pronounced in conversational speech than in read speech, we expected that CD-FWM will give greater recognition improvement in SVitchboard than in WSJ1. However, opposite result was obtained. We performed a series of analyses to investigate the possible reasons for the anomalous result, hoping that they will shed some lights for the future direction of the research area.

TABLE I
Recognition performance of CD-FWM on WSJ1 Hub2 5K evaluation task. (L is the number of phones in a word.)

Model	Word Acc.
cross-word triphones	91.53%
CD-FWM for $L \geq 6$	91.55%
CD-FWM for $L \geq 6$ + phone deletion	92.30%
CD-FWM for $L \geq 4$	91.58%
CD-FWM for $L \geq 4$ + phone deletion	92.40%

This paper is organized as follows. Explicit modeling of phone deletion using CD-FWM is described in the next section, which is followed by the experimental evaluation on SVitchboard in Section III. Several analyses are presented in Section IV to explain the anomalous result. Finally, conclusions are drawn in Section V.

II. Explicit Modeling of Phone Deletion by Context-dependent Fragmented Word Model (CD-FWM)

Phone deletion may be realized simply by adding an appropriate state transition arc across the phone to be skipped as shown in Fig. 1. In practice, if one wants to implement phone deletion in the acoustic modeling level, one has to choose a

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

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

[2]The recognition performance improvement is bigger than the one previously reported in [7] because the last phone of a word is also allowed to be skipped in the new result.

TABLE II
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^d ah-n^s^ih^d+er n^s^ih^d-er+?

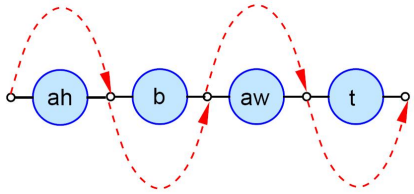

Fig. 1. An example of adding skip arcs to allow phone deletions.

linguistic unit larger than a phone for its implementation. Inspired by the work in [8], [9], we proposed *context-dependent fragmented (whole) word models* (CD-FWM) to implement phone deletions, which are constructed from well-trained tied-state cross-word triphones [7].

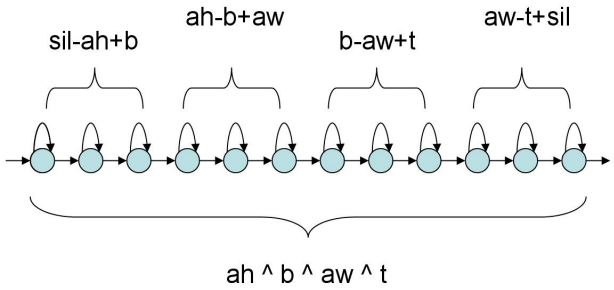

Fig. 2. An example of the construction of a context-independent word model from word-internal triphones for the word "about".

A. 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 [9], we propose the *context-dependent fragmented word models (CD-FWM)* and split a word into three or four segments as follows, depending on the length L of the word, which is defined as the number of phones in its canonical pronunciation.

- $L \leq 3$: 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 segments consisting of a single phone. Table II gives an example of a 3-segment CD-FWM for the word "about".
- $L \geq 6$: the word is split into 4 segments with the first two and the last segment consisting of a single phone. Table II gives an example of a 4-segment CD-FWM for the word "consider".

The idea is to greatly reduce the number of of possible context-dependent units by shielding the center segment from cross-word contexts in the CD-FWM. 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 segment of a 4-segment CD-FWM is just an original cross-word triphone; the remaining 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 phones and V is the size of the vocabulary, instead of $O(V^3)$ if "tri-words" are used.

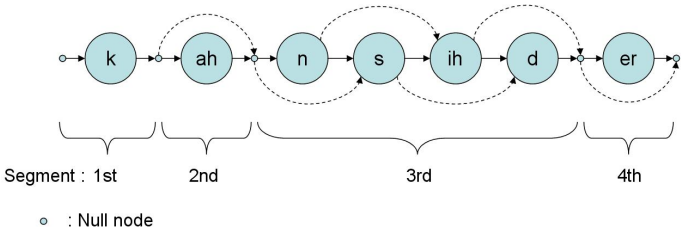

Fig. 3. An example of adding skip arcs to allow phone deletions in the actual implementation of context-dependent fragmented word models (CD-FWM) for the word "consider".

B. Remarks

We should not allow *all* phones to be skipped, otherwise a word virtually becomes a *null* word. In our current work, the following constraints are imposed on the addition of skip arcs to a CD-FWM:

- The first phone of a word is always kept. As a matter of fact, it is found that the onset portion of a syllable is usually preserved regardless of speaking conditions [1], [2]. However, in our work, we only preserve the first phone of the first syllable of a word. That is, the first

segment of the CD-FWMs for any word with $L \geq 4$ is not skipped.
- In practice, we implement CD-FWM using the HTK toolkit. HTK does not allow two successive tee-models. It means that we may not allow two successive phonetic units — conventional triphones, additional CD phones, or SWUs — in our CD-FWM to be skipped. As a result, all phones in an SWU in the center segment of a CD-FWM are *not* allowed to be skipped simultaneously. In this work, we take a simple approach and do not allow two successive phones to be skipped in an SWU.

Fig. 3 shows an example of how skip arcs are added to the CD-FWM of the word "consider".

TABLE III
INFORMATION OF VARIOUS SVITCHBOARD 500-WORD DATA SETS.

Set	#Speakers	#Utterances	#Word Tokens	Duration
train	324	13,597	51,324	3.69 hrs
dev	107	4,871	18,075	1.32 hrs
eval	107	5,202	20,021	1.43 hrs

III. EXPERIMENTAL EVALUATION ON SVITCHBOARD

The effectiveness of modeling phone deletion in conversational speech using the proposed context-dependent fragmented word models (CD-FWM) was evaluated on the SVitchboard 500-word task.

A. Speech Corpora

SVitchboard (SVB) [10] is a conversational telephone speech corpus that is defined using subsets of the Switchboard-1 corpus [3]. It defines several small vocabulary data sets ranging from 10 to 500 words, of which each task has a completely closed vocabulary. Each data set is further divided into five partitions so that they can be used for training, development, and evaluation. The speakers of one partition do not overlap with the speakers in any other partitions. In this paper, one of the SVitchboard 500-word tasks was used. The training, development, and evaluation sets are described as follows:

- Training set: Partitions A, B, and C were used as the training data. There are totally 13,597 utterances from 324 speakers. The duration of speech in this set is 3.69 hours in total.
- Development set: Partition D was used as the development data. It consists of 4,871 utterances from 107 speakers. The duration of speech in this set is 1.32 hours in total.
- Evaluation set : Partition E was used as the testing data. It consists of 5,202 utterances from 107 speakers. The duration of speech in this set is 1.43 hours in total.

A summary of these data sets is shown in Table III.

B. Experimental Setup

The following setup was used in the conversational speech experiments:

- Feature Extraction: The 39-dimensional Perceptual Linear Prediction (PLP) vectors were extracted at every 10ms over a window of 25ms. Each PLP vector consists of 12 PLP coefficients and the normalized log energy as well as their first- and second-order derivatives.
- Dictionary: The lexicon provided by the Switchboard Transcription Project [11] was used. The number of base phones was originally 42 but it was reduced to 39 by converting [ax] to [ah], [el] to [ah l], and [en] to [ah n].
- Language Model: A bigram-backoff language model was constructed using the language modeling toolkit SRILM [12]. Only the training data set was used to train the LM.
- Decoding: Recognition was performed using the HTK toolkit [13] with a beam search threshold of 200.

C. Training of the Baseline Cross-word Triphone Models

The baseline triphone model consists of 62,402 virtual triphones and 4,558 real triphones based on 39 base phones. Each triphone model is a strictly left-to-right 3-state continuous-density hidden Markov model, with a Gaussian mixture density of at most 16 components per state, and there are totally 660 tied states. The model size was chosen to maximize the recognition accuracy of the development set. In addition, there are a 1-state short pause model and a 3-state silence model.

D. 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 phonetic representation was replaced by the corresponding CD-FWM segments. Note that the number of segments in the CD-FWM of a word depends on its length as described in Section II-A. 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 IV.

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 essentially 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 phones according to the rules described in Section II-A.

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

TABLE IV
RECOGNITION PERFORMANCE ON THE SVITCHBOARD 500-WORD TASK. THE NUMBERS IN THE BRACKETS ARE THE NUMBER OF VIRTUAL UNITS. (SWU = SUB-WORD UNITS)

Model	#CD Phones	#SWUs	#Skip Arcs	Word Acc.
cross-word triphones	4,558 (62,402)	0	0	44.17%
CD-FWM for $L \geq 6$	4,631 (65,599)	79 (79)	0	44.18%
CD-FWM for $L \geq 6$ + phone deletion	4,631 (65,599)	79 (79)	567 (3,513)	44.23%
CD-FWM for $L \geq 4$	4,908 (78,679)	249 (250)	0	44.33%
CD-FWM for $L \geq 4$ + phone deletion	4,908 (78,679)	249 (250)	1,549 (10,427)	44.43%

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 baseline cross-word triphones that derive them, due to the SWUs (which are represented by the center segments in the CD-FWMs), after re-training the acoustic models that involve those center segments (e.g., ?-ah+bˆaw in Table II) 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 phones). The state distributions might also be different after re-training.

E. Results

The recognition performance of the cross-word triphone baseline and various CD-FWM systems are shown in Table IV.

From Table IV, we are disappointed to find that the performance of CD-FWM with phone deletion on SVitchboard is very different from its performance on the WSJ1 task: on WSJ1, it obtains a 10.3% performance gain when compared with the cross-word triphone baseline; on SVitchboard, it gives only a small recognition improvement (absolute 0.1%) which is statistically insignificant. The result is somewhat unexpected and seems to contradict with the common belief that the phone deletion phenomenon is more pronounced in conversational speech than in read speech.

IV. ANALYSIS AND DISCUSSION

In this section, we further investigate some plausible reasons for the seemingly anomalous result in two directions: coverage of long words in the SVitchboard corpus and the confusion between word models induced by the addition of skip arcs.

TABLE V
DISTRIBUTION OF WORD TOKENS OF DIFFERENT LENGTHS IN THE TEST SETS OF WSJ1 AND THE SVITCHBOARD 500-WORD TASK.

Word Length	Hub2 Eval Set	SVB-500 Test Set
$L \geq 6$	942 (26%)	708 (3.5%)
$L \geq 4$	1,817 (50%)	4,130 (20.6%)
$L \geq 1$	3,647 (100%)	20,021 (100%)

1) Analysis I: Frequency Distribution of Words of Different Lengths: In [2], it has been shown that words differ greatly in their frequency of occurrence in written English and spoken English. In conversational speech, the most common words occur far more frequently than the least ones, and most of them are short words with few phones. Thus, we compare the frequency distributions of word tokens[3] of different lengths in WSJ1 and SVB 500-word task. The result is summarized in Table V and the details are plotted in Fig. 4. It can be seen that only 3.5% of words in the test set of the SVB 500-word task have at least 6 phones while the similar figure in WSJ1 Hub2 eval set is 26%. If we include words with 4 or 5 phones, the figures are 20.6% and 50% respectively.

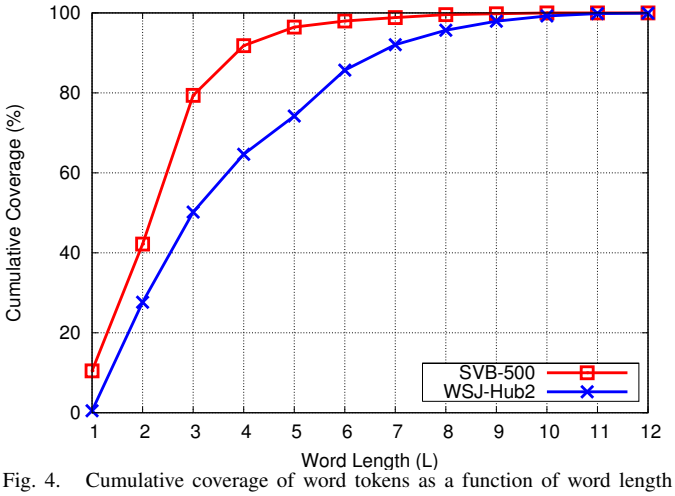

Fig. 4. Cumulative coverage of word tokens as a function of word length in the WSJ1 Hub2 eval set and the SVB 500-word test set.

Now a detailed examination of the phone deletion modeling result in Table I of the WSJ1 read speech experiment shows that most of the recognition gain comes from modeling phone deletion in words of $L \geq 6$, whereas further modeling phone deletion in words of $L = 4$ or 5 gives a modest additional gain of only 0.1%. If the same behaviour of our phone deletion modeling method is assumed to carry over to the conversational SVB 500-word task, then since the task contains so few words of $L \geq 6$, the improvement in recognition could be negligible.

2) Analysis II: Confusions Induced by Phone Deletion Modeling: Let us first denote the CD-FWM *without* phone deletion skip arcs as Model NP and CD-FWM *with* phone deletion skip arcs as Model P. We then investigate the confusion between the two models, NP and P, by doing the following analysis:

- For each test utterance, the recognized sentence produced by each model, NP or P, is aligned with the reference

[3]During the counting of word tokens, if a word appears multiple times, it is counted multiple times.

TABLE VI
CONFUSION BETWEEN CD-FWM WITHOUT PHONE DELETION MODELING AND CD-FWM WITH PHONE DELETION MODELING.

CD-FWM Without Phone Deletion	CD-FWM With Phone Deletion	
	Correct	Wrong
Correct	9,961	309
Wrong	342	9,407

transcription.

- Thus, for each word in the reference transcriptions, we may know if each of the two models recognizes it correctly: wrong recognitions are caused by substitution or deletion errors; insertion errors are not taken into account in this analysis.
- Each word in the reference transcriptions may be classified into one of the following four categories:
  1) correctly recognized by both Model NP and P.
  2) correctly recognized by Model NP but wrongly recognized by Model P.
  3) wrongly recognized by Model NP but correctly recognized by Model P.
  4) wrongly recognized by both Model NP and P.

From the numbers in Table VI, we can see that although Model P successfully "saves" 342 words that are wrongly recognized by Model NP, it misrecognizes 309 words that are correctly recognized by Model NP. As a result, in the SVitchboard 500-word task, phone deletion modeling using CD-FWM only achieves a net gain of an additional 33 correctly recognized words when compared to the system without phone deletion modeling. When also taking into account the result of Analysis I above, we believe that short words with 4 or 5 phones produce substantial confusions when some of their phones are deleted. For instance, the word "USE" may be incorrectly recognized by CD-FWM as "USED" if the last phone [d] is allowed to be deleted.

V. CONCLUSIONS

In this paper, we extend our study of explicit phone deletion modeling using *context-dependent fragmented word models* (CD-FWM) on conversational speech. Although the phone deletion phenomenon is more pronounced in conversational speech, we did not obtain statistically significant performance improvement when we used CD-FWM with phone deletion modeling on the SVitchboard 500-word recognition task. This comes at great surprise since we had been able to obtain more than 10% WER reduction in the WSJ1 read speech task before by using CD-FWM with phone deletion modeling. A closer look at the relationship between the recognition results and the distribution of the number of phones in words shows that our current CD-FWM obtains most of the recognition gain from modeling phone deletion in words with at least six phones. Moreover, a breakdown of the number of words derived from the recognized outputs from CD-FWM with and without phone deletion modeling indicates that when phones are deleted from the short words in the SVitchboard 500-word task, they are readily confused with other words in its lexicon. Hence, the analyses suggest that future work in modeling phone deletion by CD-FWM concentrate on reducing confusions when skip arcs are added to short words with fewer than six phones.

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. Ainsworth and S. Greenberg, Eds. Keele University, UK, 1996, pp. 1–8.

[2] ——, "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 *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Mar. 1992, pp. 517–520.

[4] A. Sethy, B. Ramabhadran, and S. Narayanana, "Improvements in English ASR for the MALACH project using syllable-centric models," in *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop*, 2003.

[5] A. Hämäläinen, L. Bosch, and L. Boves, "Construction and analysis of multiple paths in syllable models," in *Proceedings of Interspeech*, 2007, pp. 882–885.

[6] 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 *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2001.

[7] T. Ko and B. Mak, "Improving speech recognition by explicit modeling of phone deletions," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Dallas, Texas, USA, March 2010, pp. 4858–4861. [Online]. Available: PDF/icassp2010-pd.pdf

[8] A. Sethy and S. Narayanana, "Split-lexicon based hierarchical recognition of speech using syllable and word level acoustic units," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. I, 2003, pp. 772–775.

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

[10] S. King, C. Bartels, and J. Bilmes, "SVitchboard 1: Small vocabulary tasks from switchboard 1," in *Proceedings of Interspeech*, 2005.

[11] S. Greenberg, "The switchboard transcription project," in *Continuous Speech Recognition Summer Research Workshop Technical Report Series.* Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD, 1997.

[12] A. Stolcke, "Srilm an extensible language modeling toolkit," in *Proceedings of the International Conference on Spoken Language Processing*, 2002.

[13] S. Young *et al.*, *The HTK Book (Version 3.4)*. University of Cambridge, 2006.