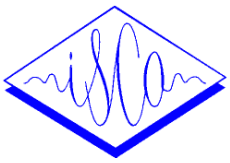

ISCA & IEEE Workshop on Spontaneous
Speech Processing and Recognition
Tokyo Institute of Technology, Tokyo, Japan
April 13–16, 2003

# FILLED-PAUSE MODELING FOR MEDICAL TRANSCRIPTIONS

*Hauke Schramm, Xavier L. Aubert, Carsten Meyer, Jochen Peters*

Philips Research Laboratories
Weisshausstrasse 2, 52066 Aachen, Germany
{hauke.schramm, xavier.aubert, carsten.meyer, jochen.peters}@philips.com

## ABSTRACT

We present our recent progress in filled pause (FP) modeling for a highly spontaneous medical transcription task. Our studies confirm that FP modeling is an important topic for spontaneous speech applications, which must be explicitly addressed in acoustic, lexical, and language modeling. We provide a framework for data-driven lexical modeling of FP acoustic variability with respect to phonemic realization and duration. By using a number of properly weighted FP pronunciation variants of variable lengths and applying specific acoustic models for FP, we achieved an 8% relative reduction of the word error rate.

We also tested different approaches for handling FP in the language model and integrating FP into the decoder. Best results with respect to both perplexity and word error rate have been achieved by predicting FP probabilistically and removing it from the language model history. This approach reduces the perplexity by 4% and provides a further gain in word accuracy.

## 1. INTRODUCTION

It is well known that speech disfluencies, especially prevalent in spontaneous speech, are an important source of confusion for ASR. A substantial part of disfluencies is formed by filled pauses (FP) like “UH” or “UHM”, which therefore must be taken into account when building spontaneous speech applications. In previous work [7] the usage of acoustic features for FP detection has been suggested. Other researchers have proposed to explicitly model the most frequent disfluency types in the language model (LM) [12]. In this approach disfluencies have been predicted probabilistically and removed from the LM history. However, on Switchboard, the application of this approach did not have a significant impact on recognition accuracy. The usage of specific context-dependent acoustic models for FP has been introduced in [3] for the Broadcast News task, where FP have also been explicitly represented in the LM. Significant improvements on the Broadcast News task through explicit FP modeling in acoustic and language model have been reported in [6]. In this study particularly long lexical pronunciations for FP were introduced in order to reduce the lexical confusability with other words. The impact of modeling representations of non-speech events in both acoustic and language model was also investigated in [8]. Here it was shown for telephone-based natural language understanding tasks that explicit FP modeling is favourable especially for real-time conditions.

In this paper, we study a number of FP modeling techniques on a highly spontaneous medical transcription task.

- First, an improved acoustic and lexical model representation for FP events is achieved by (1) applying a data-driven weighting and selection scheme of proper FP pronunciation variants with various lenghts and (2) using specific acoustic models for FP.
- Second, different ways for predicting FP events as well as handling FP in the conditioning LM history are explored, to this end extending our “standard” search algorithm.
- Third, a FP-specific error analysis is provided, highlighting the effect of LM-induced errors by misrecognized FP.

In Section 2 we describe the task. The experimental setup and our acoustic and lexical modeling approach is explained in section 3. In section 4 we present several approaches for handling FP in the language model, with decoding experiments reported in section 5. Section 6 summarizes our findings.

## 2. MEDICAL TRANSCRIPTION TASK

For our experiments we use an inhouse data collection of real-life recordings of medical reports, spontaneously spoken over long distance telephone lines by various speakers from all over the US. This database contains a variety of speaking styles, accents and speaking rates and a large degree of spontaneous speech effects like FP, partial words, repetitions and restarts. The acoustic training corpus consists of about 130h of data (295 speakers, 1.1M words), where FP and noise events are annotated. The development corpus (DEV set) consists of 5.0h of speech (11 speakers, 38.0k words), the evaluation corpus (EVAL set) of 3.3h of speech (11 speakers, 26.5k words).

## 3. ACOUSTIC AND LEXICAL MODELING OF FILLED PAUSES

### 3.1. Modeling aspects

In this section we present our approach to improve the handling of FP in the acoustic and lexical model. By extending the lexical pronunciation model of FP we try to capture the observed variability of FP acoustic with respect to phonemic realization and duration. Similar to [6], we introduce a number of pronunciation variants for FP, incorporating alternate regular and artificially lengthened (e.g. “ah-ah-ah-m”) phoneme sequences. We however do not fix the number of phonemes for the FP pronunciations. The resulting variability and increased confusability with regular words is controlled by using unigram prior probabilities for the different FP pronunciations. Furthermore, we use FP-specific phoneme symbols as presented in [3].

### 3.2. Experiments

All experiments presented in this paper have been carried out using the Philips Research large vocabulary continuous speech recognizer which has been previously described in [1, 2]. Results are

reported for the first decoding pass without any acoustic or language model adaptation.

The baseline system performs a one-pass trigram decoding with a 75K vocabulary, the alternative pronunciations accounting for about 20% of all entries. The score contributions of simultaneously active pronunciations of the same word are summed up during decoding, which has been shown to be beneficial both in terms of accuracy and efficiency [9]. Standard language and phoneme look-ahead techniques are conservatively applied to provide a less than 10 times real-time running factor[1]. In the experiments of this section FP is handled like any other word by the LM, i.e. it is predicted probabilistically from trigram context and may also appear in the history itself. Note that FP events are removed from the spoken and recognized texts prior to word error rate computation. This means that (1) deletions and insertions of FP are not counted as errors, (2) the substitution of a word by a FP is counted as a deletion while the substitution of a FP by a word is counted as an insertion error.

For training of our baseline system we used transcriptions with manually labelled FP pronunciation variants to estimate standard context-dependent triphone models. In decoding, only a single pronunciation of FP was used, corresponding to the most frequent variant in the manual FP training transcripts. On the DEV set, this baseline system achieved a word error rate of 24.4% (Table 1, ID 1). Using the four most frequent FP pronunciation variants in decoding, we observed a small improvement of 0.3% absolute (Table 1, ID 2). Here, equal pronunciation variant weights have been applied for the four lexical representations of FP.

In the next step the labeling of FP pronunciations in the training transcripts was done automatically by performing a standard Viterbi alignment with free choice between the different FP pronunciation variants. This technique allows to train arbitrary sets of FP pronunciations by ensuring their consistent labelling in the training transcripts and enables the estimation of respective FP unigram prior probabilities. In training, we used the $N$ most frequent FP pronunciations as determined from an initial labeling pass with the complete set of FP pronunciations. The restricted set of FP variants was in turn used (1) in a second alignment pass to generate the final training transcript and (2) in the training and decoding lexicons. The effect of an automatic FP variant selection for $N = 4$ variants, in combination with the application of FP-specific phoneme symbols, is demonstrated in ID 3 (Table 1). Using equal prior probabilities for the FP variants in decoding, we obtained a WER of 23.4%. Introducing phonetic transcriptions of various lengths for each regular FP pronunciation and again restricting training and recognition to the four most frequent variants, a gain of 0.4% absolute was obtained (Table 1, ID 4). Using the 16 most frequent variants with equal prior probabilities in decoding, we did not observe any gain (ID 5). However applying unigram priors instead was beneficial to give a further 0.3% absolute WER reduction (ID 6). Finally, with 64 FP variants, we obtained a WER of 22.5% on the DEV set (ID 7 in Table 1).

### 3.3. Error analysis

Accounting for FP in both acoustic and lexical modeling leads to a significant improvement of the overall word error rate. Generally, several effects might account for this gain:

- reduced confusability between words and FP,
- more accurate LM histories, due to less FP insertion, deletion and substitution errors, reducing the error propagation effects (see example below),
- improved acoustic models for regular (i.e. non-FP) speech due to better segmentation and more specific mixtures.

[1]There is no evidence, so far, that this setup introduces significant search errors.

| ID | SM | LP | PW | #PV | WER(%) |
|---|---|---|---|---|---|
| 1 | no | no | no | 1 | 24.4 (ref.) |
| 2 | no | no | no | 4 | 24.1 (-1.2%) |
| 3 | yes | no | no | 4 | 23.4 (-4.1%) |
| 4 | yes | yes | no | 4 | 23.0 (-5.7%) |
| 5 | yes | yes | no | 16 | 23.0 (-5.7%) |
| 6 | yes | yes | yes | 16 | 22.7 (-7.0%) |
| 7 | yes | yes | yes | 64 | 22.5 (-7.8%) |

**Table 1**. *Results achieved on DEV set with different setups for the integration of FP in acoustic and lexical model. SM: FP **s**pecific acoustic **m**odels, LP: **l**ong FP **p**ronunciations, PW: **p**ronunciation unigram **w**eights applied for FP variants, #PV: number of FP pronunciation variants in decoding lexicon.*

To identify the primary effects responsible for the observed improvements, we performed a *FP specific error analysis* on the "full" transcripts (including FP).

We distinguish between four different types of FP specific errors: *deletions* and *insertions* of FP, replacement of a word by a FP (what we call FP "*substitution_insertion*" error) and replacement of a FP by a word (FP "*substitution_deletion*" error). In Table 2 examples are given for these error types. These FP related errors

| Error Type | Spoken | Recognized |
|---|---|---|
| FP insertion | has had several | has had FP several |
| FP deletion | the FP aortic valve | the aortic valve |
| FP substitution_insertion | have a limited | have FP limited |
| FP substitution_deletion | abscess FP she is | abscess that she is |

**Table 2**. *Examples for the different FP specific error types*

have a direct effect on the word error rate but may in addition influence the prediction of successor words consecutive to a perturbed LM history and acoustic segmentation. A real-life example for such an *error propagation* effect is given below. The replacement of FP ("UH") by a word ("on") significantly changes the trigram LM history for the prediction of the word "of", thereby causing its deletion.

```
Spoken : PERIOD UH Examination of the chest
Recogn.: PERIOD On examination -- the chest
```

Table 3 presents the FP specific error counts for the baseline system (ID 1 in Table 1 and 3) and for our best system with 64 FP pronunciation variants (ID 7 in Table 1 and 3). In addition, we give the number of errors of regular words (without FP), labelled "WEC", and the *recall* and *precision* rate for FP. The recall specifies the rate of spoken FP that have been recognized correctly while the precision is defined as the rate of recognized FP that have actually been spoken. The test set contains a total number of 2852 spoken FP. Comparing the first two rows in Table 3 we observe that the improved acoustic and lexical modeling of FP lead to a significant increase of the recall rate. On the other hand, the precision is strongly reduced, corresponding to a large increase of insertion and substitution_insertion errors. Thus, apart from better matching spoken FP, the advanced FP modeling leads to increased confusability with "regular" words. Apparently the latter effect is even stronger, since we observe a strong increase in

| ID | WEC | Ins | Del | Sub_I | Sub_D | Recall | Prec. |
|---|---|---|---|---|---|---|---|
| 1 | 9275 | 123 | 546 | 221 | 779 | 53.5% | 81.6% |
| 7 | 8558 | 833 | 105 | 1381 | 180 | 90.0% | 53.7% |
| 8 | 10289 | 0 | 793 | 0 | 2059 | 0% | - |

**Table 3**. *FP specific insertion (Ins), deletion (Del), substitution_insertion (Sub_I) and substitution_deletion (Sub_D) errors observed for different levels of FP modeling. "WEC" is the total word error count ignoring FP.*

substitution_insertion errors which exceeds the decrease in substitution_deletion errors by 561 counts. To explain the overall error reduction of 717 words, we hypothesize (1) better acoustic modeling of regular (i.e. non-FP) speech as side-effect of improved FP modeling and (2) error propagation effects as explained above. We infer that the replacement of FP by a word (substitution_deletion) is more harmful to the LM history than the replacement of a word by FP (substitution_insertion), since FP is much less specific to the LM than regular words. To check this hypothesis, we performed a contrast experiment, using our best system (Table 3, ID 7), however removing FP from the decoding lexicon (ID 8). As can be seen in Table 3, in 2059 (out of 2852) cases a spoken FP was recognized as a regular word (instead of being deleted). Contrasting these 2059 directly FP-related errors to the 1381+180=1561 directly FP-related errors of our best system (Table 3, ID 7), we see that the increase of the total error rate in the contrast experiment by 10289-8558=1731 can only partly be explained by direct FP-related effects. The larger part of the additional word errors were introduced as a *side effect* of misrecognized FP, presumably due to the LM history effect explained above. This demonstrates that a misrecognized FP causes additional LM-induced errors.

## 4. HANDLING OF FILLED PAUSES IN LM

The optimal handling of disfluencies in the language model (LM) is not straightforward [11, 12]. Concentrating on FP only, we may (1.) treat FP as a regular word which is predicted by the LM and which conditions following words (as has been done in Section 3), or (2.) use the LM for prediction of both words and FP but discard all FP from the conditioning histories, or (3.) use a fixed, context-independent probability for FP. In approach 3. like in 2., the LM probabilities are always conditioned on non-disfluent text, i.e. words are conditioned on preceding words skipping FP. Table 4 summarizes these approaches. In approach 1. the influence

| | $\text{Prob}(W \mid U, V, \text{FP})$ | $\text{Prob}(\text{FP} \mid U, V)$ |
|---|---|---|
| 1. | $p_{\text{LM}}(W \mid V, \text{FP})$ | $p_{\text{LM}}(\text{FP} \mid U, V)$ |
| 2. | $p_{\text{LM}}(W \mid U, V)$ | $p_{\text{LM}}(\text{FP} \mid U, V)$ |
| 3. | $p_{\text{LM}}(W \mid U, V)$ | $p_{\text{fixed}}(\text{FP})$ |

**Table 4**. *Usage of trigram probab. $p_{\text{LM}}$ for approach 1.-3.*

of FP on the next spoken words is taken into account. However, in this case the M-Gram linguistic context is reduced or even lost when several FP are spoken in sequence. Approach 2. preserves the full linguistic context but neglects any predictive influence of FP on following words. In the following, we experimentally evaluate the three approaches.

### 4.1. Training data

Two corpora have been used for estimating the LMs (see Table 5). The first one contains transcripts of spontaneously spoken medical reports with annotated FP. The second training corpus comprises formatted medical reports which were transformed into a spoken-like form, e.g. by mapping numbers and some abbreviations to the recognizer's vocabulary. Some characteristics of the spontaneous data (such as often non-dictated punctuation) are approximated by stochastic mappings. However, most spontaneous effects and all disfluencies are absent. For training of the LMs *including* FP, this report corpus was stochastically enriched with FP: Considering FP as hidden events in the reports we randomly inserted them based on the a-posteriori probabilities in the given word contexts. These probabilities were derived from a bigram trained on the FP-annotated spontaneous training data.

| Corpus | Size | FP rate | OOV rate |
|---|---|---|---|
| Spontaneous training | 1314 k | 8.2 % | 0.45 % |
| Mapped reports | 1071 M | 7.9 % | 0.31 % |
| Spont. development | 81 k | 6.3 % | 0.23 % |

**Table 5**. *Sizes of text corpora including FP.*

### 4.2. Language model training

On the two corpora, we trained trigrams both with and without FP. The trigrams from the huge report corpus were also pruned to a reasonable size using the method described in [4].

Table 6 summarizes the LM sizes. Note that FP-free trigrams distinguish more M-Grams than trigrams with FP. This is due to a clear bias of FP to appear before or after certain words.[2] Here, observed word transitions are "destroyed" by interposed FP. Our best LMs use merged count statistics (with and without FP). The reduced FP rate is compensated by "marginal adaptation" [5]. Finally, the LMs from both corpora were interpolated with weights optimized on our development corpus.

| Corpus | Pruned | Count statistics | | |
|---|---|---|---|---|
| | | with FP | FP-free | merged |
| Spont. | no | 990 k | 1001 k | 1150 k |
| Reports | no | 98 M | 107 M | 117 M |
| Reports | yes | 17 M | 18 M | 18 M |

**Table 6**. *Number of observed uni- + bi- + trigrams.*

### 4.3. Perplexities

Table 7 shows perplexities on the development corpus. Comparing approaches 1. and 2. we see that the removal of FP from the word history reduces perplexity by 4%. Contrary to observations in [12, 8], FP are – on average – worse predictors for following words than the FP-free context. We also observe that the merged count statistics improve LM performance by 2–3% due to the "recovery" of word M-Grams which are unobserved in data with FP.

### 4.4. Prediction of FP

Perplexities are almost identical for the best approach (2. with merged count statistics) and for approach 3. One fundamental difference, however, is not reflected by the overall perplexity: Probabilities for FP are *context-dependent* for approach 1. and 2., but a *fixed* value is used for approach 3. In a contrast experiment we measured the average probability of FP for (1) all histories followed by a word ($\neq$ FP), and (2) all histories followed by FP. Table 8 shows that FP probabilities are reduced by almost 50% for

[2] The effect disappears if FP is randomly inserted into an FP-free corpus in a *context-independent* manner.

| Approach | Pruned | Count statistics | |
|---|---|---|---|
| | | with FP | merged |
| 1. | no | $61.4 \pm 1.4$ | $60.3 \pm 1.3$ |
| 2. | no | $59.2 \pm 1.3$ | $57.5 \pm 1.2$ |
| 3. | no | $57.9 \pm 1.2$ | |
| 1. | yes | $63.7 \pm 1.4$ | $62.6 \pm 1.4$ |
| 2. | yes | $61.5 \pm 1.4$ | $59.9 \pm 1.3$ |
| 3. | yes | $60.0 \pm 1.3$ | |

**Table 7**. *Perplexities and error bars (95% confidence) for interpolated trigrams. For approach 3. we use an FP-free trigram and* $p_{\text{fixed}}(FP) = 0.08 =$ *FP rate in training.*

approach 1. and 2. in "word positions" as compared to "FP positions". Approach 3. naturally lacks this discriminative power.

| Approach | Pruned | word positions | FP positions |
|---|---|---|---|
| 1. | no | $18.8 \pm 0.2$ | $9.8 \pm 0.2$ |
| 2. | no | $20.9 \pm 0.2$ | $11.4 \pm 0.2$ |
| 3. | no | $12.5 \pm 0.0$ | $12.5 \pm 0.0$ |

**Table 8**. *Inverse FP probabilities (geom.mean). Approach 1. and 2. uses merged counts.*

## 5. HANDLING OF FILLED PAUSES IN DECODING

In this section we evaluate the LM approaches from section 4 in decoding experiments. Some modifications have been introduced into the search algorithm to implement the distinct ways of handling FP, sketched in Table 4. This mainly concerns the integration of the LM probabilities in the word ending recombination scheme. Figure 1 describes the two scenarios. In the upper part, a FP is

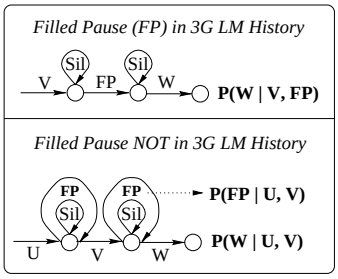


**Fig. 1**. *Handling of FP with a 3-Gram Language Model*

handled like any other word and may appear in the LM history. The second scheme is illustrated in the lower part of the figure. Here, FP is deliberately excluded from the LM history although it is conditioned on the previously spoken words.

We performed experiments on predicting FP for three different history lengths, namely, using a unigram, bigram or trigram stochastic model. Additionally, a FP specific penalty has been considered. The main results are summarized in Table 9, giving the final word error rates obtained on the EVAL set[3]. Here, the best acoustic and lexical models obtained in Section 3 have been combined with different ways of handling FP in the LM. The best results are obtained when FP are excluded from the trigram histories though predicted with trigram probabilities. In all four cases, the rate of deletions and insertions is equal to 13.4% which means that the gain is obtained through a reduction of the substitutions from 16.65% to 16.0% representing a 4% relative improvement.

[3]Please note that, due to some outlier speakers, the word error rate for the EVAL set is higher than for the DEV set.

| Case | LM History | FP Prediction | WER % | Improvement |
|---|---|---|---|---|
| Base | FP included | Std Trigram | 30,07% | Ref. |
| 1 | FP excluded | Unigram | 29,95% | - |
| 2 | FP excluded | Bigram | 29,84% | -0.8% |
| 3 | FP excluded | Trigram | 29,41% | -2.2% |

**Table 9**. *Word Error Rates (WER) achieved on the EVAL set.*

## 6. SUMMARY

In this paper we presented various approaches to improve ASR performance on a highly spontaneous medical transcription task by introducing sophisticated FP modeling techniques into our system. We suggest using unigram prior probabilities to control the variability in FP modeling in a data-driven way. Applying FP pronunciation variants with a variable number of FP-specific phonemes, we achieved a reduction of the word error rate by 8% relative. A FP-specific error analysis showed that misrecognized FP cause additional LM-induced error propagation.

In the second part, we tested several approaches for handling FP in the language model and the decoder. Modest additional gains have been obtained by applying LM prediction for both words and FP but discarding all FP from the conditioning histories.

## 7. REFERENCES


[1] X. Aubert, "One Pass Cross Word Decoding For Large Vocabularies Based On A Lexical Tree Search Organization", Proc. Eurospeech 1999, pp. 1559-1562, Budapest, Hungary.

[2] X. Aubert, R. Blasig, "Combined Acoustic and Linguistic Look-Ahead for One-Pass Time-Synchronous Decoding", in Proc. ICSLP 2000, Vol. 3, pp. 802-805, Bejing, China.

[3] J.L. Gauvain, G. Adda, L. Lamel, M. Adda-Decker, "Transcribing Broadcast News: The LIMSI Nov96 Hub4 System", in Proc. DARPA Speech Recognition Workshop, 1997.

[4] R. Kneser, "Statistical language modeling using a variable context length", in Proc. ICSLP 1996, Vol. 1, pp. 494–497.

[5] R. Kneser, J. Peters, and D. Klakow, "Language model adaptation using dynamic marginals" in Proc. Eurospeech 1997, Vol. 4, pp. 1971–1974.

[6] D. Liu, L. Nguyen, S. Matsoukas, J. Davenport, F. Kubala, R. Schwartz, "Improvements in spontaneous speech recognition", in Proc. DARPA Speech Recognition Workshop, 1999.

[7] D. O'Shaughnessy, "Recognition of hesitations in spontaneous speech", in Proc. ICASSP 1992, Vol. 1, pp. 521 -524.

[8] R.C. Rose, G. Riccardi, "Modeling disfluency and background events in ASR for a natural language understanding task", in Proc. ICASSP 1999, Vol. 1, pp. 341 -344.

[9] H. Schramm, X. Aubert, "Efficient integration of multiple pronunciations in a large vocabulary decoder", Proc. ICASSP 2000, pp. 1659–1662, Vol. 3, Istanbul, Turkey.

[10] E. Shriberg, A. Stolcke, "Word predictability after hesitations: a corpus-based study", in Proc. ICSLP 1996, Vol. 3, pp. 1868 -1871.

[11] M. Siu; M. Ostendorf, "Modeling disfluencies in conversational speech", in Proc. ICSLP 1996, Vol. 1, pp. 386 -389.

[12] A. Stolcke, E. Shriberg, "Statistical language modeling for speech disfluencies", in Proc. ICASSP 1996, pp. 405 -408.