Modeling Filled Pauses in Medical Dictations

Sergey V. Pakhomov

University of Minnesota
190 Klaeber Court
320-16th Ave. S.E. Minneapolis, MN 55455
pakh0002@tc.umn.edu

Abstract

Filled pauses are characteristic of spontaneous speech and can present considerable problems for speech recognition by being often recognized as short words. An *um* can be recognized as *thumb* or *arm* if the recognizer's language model does not adequately represent FP's. Recognition of quasi-spontaneous speech (medical dictation) is subject to this problem as well. Results from medical dictations by 21 family practice physicians show that using an FP model trained on the corpus populated with FP's produces overall better results than a model trained on a corpus that excluded FP's or a corpus that had random FP's.

Introduction

Filled pauses (FP's), false starts, repetitions, fragments, etc. are characteristic of spontaneous speech and can present considerable problems for speech recognition. FP's are often recognized as short words of similar phonetic quality. For example, an *um* can be recognized as *thumb* or *arm* if the recognizer's language model does not adequately represent FP's. Recognition of quasi-spontaneous speech (medical dictation) is subject to this problem as well. The FP problem becomes especially pertinent where the corpora used to build language models are compiled from text with no FP's. Shriberg (1996) has shown that representing FP's in a language model helps decrease the model's perplexity. She finds that when a FP occurs at a major phrase or discourse boundary, the FP itself is the best predictor of the following lexical material; conversely, in a non-boundary context, FP's are predictable from the preceding words. Shriberg (1994) shows that the rate of disfluencies grows exponentially with the length of the sentence, and that FP's occur more often in the initial position (see also Swerts (1996)).

This paper presents a method of using bigram probabilities for extracting FP distribution from a corpus of hand-transcribed data. The resulting bigram model is used to populate another training corpus that originally had no FP's. Results from medical dictations by 21 family practice physicians show that using an FP model trained on the corpus populated with FP's produces overall better results than a model trained on a corpus that excluded FP's or a corpus that had random FP's. Recognition accuracy improves proportionately to the frequency of FP's in the speech.

1. Filled Pauses

FP's are not random events, but have a systematic distribution and well-defined functions in discourse. (Shriberg and Stolcke 1996, Shriberg 1994, Swerts 1996, Macalay and Osgood 1959, Cook 1970, Cook and Lalljee 1970, Christenfeld, et al. 1991) Cook and Lalljee (1970) make an interesting proposal that FP's may have something to do with the listener's perception of disfluent speech. They suggest that speech may be more

comprehensible when it contains filler material during hesitations by preserving continuity and that a FP may serve as a signal to draw the listeners attention to the next utterance in order for the listener not to lose the onset of the following utterance. Perhaps, from the point of view of perception, FP's are not disfluent events at all. This proposal bears directly on the domain of medical dictations, since many doctors who use old voice operated equipment train themselves to use FP's instead of silent pauses, so that the recorder wouldn't cut off the beginning of the post pause utterance.

2. Quasi-spontaneous speech

Family practice medical dictations tend to be pre-planned and follow an established SOAP format: (Subjective (informal observations), Objective (examination), Assessment (diagnosis) and Plan (treatment plan)). Despite that, doctors vary greatly in how frequently they use FP's, which agrees with Cook and Lalljee's (1970) findings of no correlation between FP use and the mode of discourse. Audience awareness may also play a role in variability. My observations provide multiple examples where the doctors address the transcriptionists directly by making editing comments and thanking them.

3. Training Corpora and FP Model

This study used three base and two derived corpora. Base corpora represent three different sets of dictations described in section 3.1. Derived corpora are variations on the base corpora conditioned in several different ways described in section 3.2.

3.1 Base

- Balanced FP training corpus (BFP-CORPUS) that has 75, 887 words of word-by-word transcription data evenly distributed between 16 talkers. This corpus was used to build a BIGRAM-FP-LM which controls the process of populating a no-FP corpus with artificial FP's.

- Unbalanced FP training corpus (UFP-CORPUS) of approximately 500,000 words of all available word-by-word transcription data from approximately 20 talkers. This corpus was used only to calculate average frequency of FP use among all available talkers.

- Finished transcriptions corpus (FT-CORPUS) of 12,978,707 words contains all available dictations and no FP's. It represents over 200 talkers of mixed gender and professional status. The corpus contains no FP's or any other types of disfluencies such as repetitions, repairs and false starts. The language in this corpus is also edited for grammar.

3.2 Derived

- CONTROLLED-FP-CORPUS is a version of the finished transcriptions corpus populated stochastically with 2,665,000 FP's based on the BIGRAM-FP-LM.

- RANDOM-FP-CORPUS-1 (normal density) is another version of the finished transcriptions corpus populated with 916,114 FP's where the insertion point was selected at random in the range between 0 and 29. The random function is based on the average frequency of FPs in the unbalanced UFP-CORPUS where an FP occurs on the average after every 15^{th} word. Another RANDOM-FP-CORPUS-2 (high density) was used to approximate the frequency of FP's in the CONTROLLED-FP-CORPUS.

4. Models

The language modeling process in this study was conducted in two stages. First, a bigram model containing bigram probabilities of FP's in the balanced BFP-COPRUS was built followed by four different trigram language models, some of which used corpora generated with the BIGRAM-FP-LM built during the first stage.

4.1 Bigram FP model

This model contains the distribution of FP's obtained by using the following formulas:

$$P(FP|w_{i-1}) = C_{w-1\ FP} / C_{w-1}$$
$$P(FP|w_{i+1}) = C_{FP\ w+1} / C_{w+1}$$

Thus, each word in a corpus to be populated with FP's becomes a potential landing site for a FP and does or does not receive one based on the probability found in the BIGRAM-FP-LM.

4.2 Trigram models

The following trigram models were built using ECRL's Transcriber language modeling tools (Valtchev, et al. 1998). Both bigram and trigram cutoffs were set to 3.

- NOFP-LM was built using the FT-CORPUS with no FP's.
- ALLFP-LM was built entirely on CONTROLLED-FP-CORPUS.
- ADAPTFP-LM was built by interpolating ALLFP-LM and NOFP-LM at 90/10 ratio. Here 90 % of the resulting ADAPTFP-LM represents the CONTROLLED-FP-CORPUS and 10% represents FT-CORPUS.
- RANDOMFP-LM-1 (normal density) was built entirely on the RANDOM-FP-CORPUS-1.
- RANDOMFP-LM-2 (high density) was built entirely on the RANDOM-FP-CORPUS-2

5. Testing Data

Testing data comes from 21 talkers selected at random and represents 3 (1-3 min) dictations for each talker. The talkers are a random mix of male and female medical doctors and practitioners who vary greatly in their use of FP's. Some use literally no FP's (but long silences instead), others use FP's almost every other word. Based on the frequency of FP use, the talkers were roughly split into a high FP user and low FP user groups. The relevance of such division will become apparent during the discussion of test results.

6. Adaptation

Test results for ALLFP-LM (63.01% avg. word accuracy) suggest that the model over represents FP's. The recognition accuracy for this model is 4.21 points higher than that of the NOFP-LM (58.8% avg. word accuracy) but lower than that of both the RANDOMFP-LM-1 (67.99% avg. word accuracy) by about 5% and RANDOMFP-LM-2 (65.87% avg. word accuracy) by about 7%. One way of decreasing the FP representation is to correct the BIGRAM-FP-LM, which proves to be computationally expensive because of having to rebuild the large training corpus with each change in BIGRAM-FP-LM. Another method is to build a NOFP-LM and an ALLFP-LM once and experiment with their relative weights through adaptation. I chose the second method because ECRL Transcriber toolkit provides an adaptation tool that achieves the goals of the first method much faster. The results show that introducing a NOFP-LM into the equation improves recognition. The difference in recognition accuracy between the ALLFP-LM and ADAPTFP-LM is on average 4.9% across all talkers in ADAPTFP-LM's favor. Separating the talkers into high FP user group and low FP user group raises ADAPTFP-LM's gain to 6.2% for high FP users and lowers it to 3.3%

for low FP users. This shows that adaptation to no-FP data is, counter-intuitively more beneficial for high FP users.

7. Results and discussion

Although a perplexity test provides a good theoretical measure of a language model, it is not always accurate in predicting the model's performance in a recognizer (Chen 1998); therefore, both perplexity and recognition accuracy were used in this study. Both were calculated using ECRL's LM Transcriber tools.

7.1 Perplexity

Perplexity tests were conducted with ECRL's LPlex tool based on the same text corpus (BFP-CORPUS) that was used to build the BIGRAM-FP-LM. Three conditions were used. Condition A used the whole corpus. Condition B used a subset of the corpus that contained high frequency FP users (FPs/Words ratio above 1.0). Condition C used the remaining subset containing data from lower frequency FP users (FPs/Words ratio below 1.0). Table 1 summarizes the results of perplexity tests at 3-gram level for the models under the three conditions.

Model	Condition A	(all)	Condition B	(high)	Condition C	(low)
	Lplex	OOV rate (%)	Lplex	OOV rate (%)	Lplex	OOV rate (%)
NOFP-LM	617.59	6.35	1618.35	6.08	287.46	6.06
ADAPTFP-LM	132.74	6.35	120.69	6.08	131.70	6.06
RANDOMFP-LM-1	138.02	6.35	140.47	6.08	125.79	6.06
RANDOMFP-LM-2	156.09	6.35	152.16	6.08	145.47	6.06
ALLFP-LM	980.67	6.35	964.48	6.08	916.53	6.06

Table 1. Perplexity measurements

The perplexity measures in Condition A show over 400 point difference between ADAPTFP-LM and NOFP-LM language models. The 363,08 increase in perplexity for ALLFP-LM model corroborates the results discussed in Section 6. Another interesting result is contained in the highlighted fields of Table 1. ADAPTFP-LM based on CONTROLLED-FP-CORPUS has lower perplexity in general. When tested on conditions B and C, ADAPTFP-LM does better on frequent FP users, whereas RANDOMFP-LM-1 does better on infrequent FP users, which is consistent with the recognition accuracy results for the two models (see Table 2).

7.2 Recognition accuracy

Recognition accuracy was obtained with ECRL's HResults tool and is summarized in Table 2.

Language Model	AvgWordAcc (High FP user)	AvgWordAcc (Low FP user)
NOFP-LM	51.40 %	67.76%
RANDOMFP-LM-1 (normal density)	66.57 %	71.46 %
RANDOMFP-LM-2 (high density)	63.35 %	69.23 %
ADAPTFP-LM	67.14 %	71.24%

Table 2. Recognition accuracy tests for LM's.

The results in Table 2 demonstrate two things. First, a FP model performs better than a clean model that has no FP representation. Second, a FP model based on populating a no-FP training corpus with FP's whose distribution was derived from a

small sample of speech data performs better than the one populated with FP's at random based solely on the frequency of FP's. The results also show that ADAPTFP-LM performs slightly better than RANDOMFP-LM-1 on high FP users. The gain becomes more pronounced towards the higher end of the FP use continuum. For example, the scores for the top four high FP users are 62.07% with RANDOMFP-LM-1 and 63.51% with ADAPTFP-LM. This difference cannot be attributed to the fact that RANDOMFP-LM-1 contains fewer FP's than ADAPTFP-LM. The word accuracy rates for RANDOMFP-LM-2 indicate that frequency of FP's in the training corpus is not responsible for the difference in performance between the RANDOM-FP-LM-1 and the ADAPTFP-LM. The frequency is roughly the same for both RANDOMFP-CORPUS-2 and CONTROLLED-FP-CORPUS, but RANDOMFP-LM-2 scores are lower than those of RANDOMFP-LM-1, which allows in absence of further evidence to attribute the difference in scores to the pattern of FP distribution, not their frequency.

Conclusion

Based on the results so far, several conclusions about FP modeling can be made:

1. Representing FP's in the training data improves both the language model's perplexity and recognition accuracy.
2. It is not absolutely necessary to have a corpus that contains naturally occurring FP's for successful recognition. FP distribution can be extrapolated from a relatively small corpus containing naturally occurring FP's to a larger clean corpus. This becomes vital in situations where the language model has to be built from "clean" text such as finished transcriptions, newspaper articles, web documents, etc.
3. If one is hard-pressed for hand transcribed data with natural FP's, a random population can be used with relatively good results.
4. FP's are quite common to both quasi-spontaneous monologue and spontaneous dialogue (medical dictation).

Research in progress

The present study leaves a number of issues to be investigated further:

1. The results for RANDOMFP-LM-1 are very close to those of ADAPTFP-LM. A statistical test is needed in order to determine if the difference is significant.
2. A systematic study of the syntactic as well as discursive contexts in which FP's are used in medical dictations. This will involve tagging a corpus of literal transcriptions for various kinds of syntactic and discourse boundaries such as clause, phrase and theme/rheme boundaries. The results of the analysis of the tagged corpus may lead to investigating which lexical items may be helpful in identifying syntactic and discourse boundaries. Although FP's may not always be lexically conditioned, lexical information may be useful in modeling FP's that occur at discourse boundaries due to co-occurrence of such boundaries and certain lexical items.
3. The present study roughly categorizes talkers according to the frequency of FP's in their speech into high FP users and low FP users. A more finely tuned categorization of talkers in respect to FP use as well as its usefulness remain to be investigated.
4. Another area of investigation will focus on the SOAP structure of medical dictations. I plan to look at relative frequency of FP use in the four parts of a medical dictation. Informal observation of data collected so far indicates that FP use is more frequent and different from other parts during the

Subjective part of a dictation. This is when the doctor uses fewer frozen expressions and the discourse is closest to a natural conversation.

Acknowledgements

I would like to thank Joan Bachenko and Michael Shonwetter, at Linguistic Technologies, Inc. and Bruce Downing at the University of Minnesota for helpful discussions and comments.

References

Chen, S., Beeferman, Rosenfeld, R. (1998). "Evaluation metrics for language models," In DARPA Broadcast News Transcription and Understanding Workshop.

Christenfeld, N, Schachter, S and Bilous, F. (1991). "Filled Pauses and Gestures: It's not coincidence," Journal of Psycholinguistic Research, Vol. 20(1).

Cook, M. (1977). "The incidence of filled pauses in relation to part of speech," Language and Speech, Vol. 14, pp.135-139.

Cook, M. and Lalljee, M. (1970). "The interpretation of pauses by the listener," Brit. J. Soc. Clin. Psy. Vol. 9, pp. 375-376.

Cook, M., Smith, J, and Lalljee, M (1977). "Filled pauses and syntactic complexity," Language and Speech, Vol. 17, pp.11-16.

Valtchev, V. Kershaw, D. and Odell, J. 1998. The truetalk transcriber book. Entropic Cambridge Research Laboratory, Cambridge, England.

Heeman, P.A. and Loken-Kim, K. and Allen, J.F. (1996). "Combining the detection and correlation of speech repairs," In Proc., ICSLP.

Lalljee, M and Cook, M. (1974). "Filled pauses and floor holding: The final test?" Semiotica, Vol. 12, pp.219-225.

Maclay, H, and Osgood, C. (1959). "Hesitation phenomena in spontaneous speech," Word, Vol.15, pp.19-44.

Shriberg, E. E. (1994). Preliminaries to a theory of speech disfluencies. Ph.D. thesis, University of California at Berkely.

Shriberg, E.E. and Stolcke, A. (1996). "Word predictability after hesitations: A corpus-based study," In Proc. ICSLP.

Shriberg, E.E. (1996). "Disfluencies in Switchboard," In Proc. ICSLP.

Shriberg, E.E. Bates, R. and Stolcke, A. (1997). "A prosody-only decision-tree model for disfluency detection" In Proc. EUROSPEECH.

Siu, M. and Ostendorf, M. (1996). "Modeling disfluencies in conversational speech," Proc. ICSLP.

Stolcke, A and Shriberg, E. (1996). "Statistical language modeling for speech disfluencies," In Proc. ICASSP.

Swerts, M, Wichmann, A and Beun, R. (1996). "Filled pauses as markers of discourse structure," Proc. ICSLP.