

Filled Pause Distribution and Modeling in Quasi-Spontaneous Speech

Sergey Pakhomov and Guergana Savova
University of Minnesota, Minneapolis, USA

Abstract

Filled pauses (FP) are characteristic of spontaneous speech and present considerable problems for speech recognition by being often recognized as short words. Recognition of quasi-spontaneous speech (medical dictation) is subject to this problem as well. An *um* can be recognized as *thumb* or *arm* if the recognizer's language model does not adequately represent FP's. Representing FP's in the training corpus improves recognition. Several techniques of seeding a training corpus with FP's were evaluated to show that a stochastic method, along with random insertion uniformly distributed around the average sentence length, yield better results compared to random insertion at other ranges. The optimal method of seeding a training corpus with FP's may be linked to clause boundaries despite the fact that an imperfect method of inserting FP's at clause boundaries used in this study failed.

1. Introduction.

Filled Pauses are not random, but have a systematic distribution and well-defined functions in discourse. [1,2,3,5,8,9,16]. Cook and Lalljee [4] 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 an 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.

Filled pauses, false starts, repetitions, fragments, etc. are characteristic of spontaneous speech and present considerable problems for speech recognition. FP's are often recognized as short words of similar phonetic quality. Recognition of quasi-spontaneous speech (medical dictation) is subject to this problem as well. For example, an *um* can be recognized as *thumb* or *arm* if the recognizer's language model does not adequately represent FP's. The FP problem becomes especially pertinent where the corpora used to build language models are compiled from text with no FP's. Shriberg [12] has shown that representing FP's in a language model helps decrease the model's perplexity. She finds that when an 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 [10] 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 [16]).

In a previous study, Pakhomov [9] shows that a language model based on a training corpus populated with FP's stochastically significantly improves recognition over the language model that contains no FP's. The improvement over the model that contains FP's inserted into the training corpus at random with insertion points uniformly distributed around every 15th word (the average frequency of FP's) is very slight.

In this paper we will present an additional method of seeding training corpora with FP's where the insertion points are clustered around clause boundaries with subsequent language model generation and several ways of evaluating the recognition results. We will show that using recognition accuracy as the only gauge to determine the goodness of an FP model is not sufficient due to FP overrepresentation effects. We will also show advantages and disadvantages of populating training corpora with FP's at random

2. Quasi-spontaneous speech

The term quasi-spontaneous speech reflects the fact that medical dictations used for analysis in this study are very different from unprepared monologues as well as read text and tend to retain features of both. 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)). The Subjective part tends to resemble unrehearsed monologues where as the rest of the dictation is more like read speech. Besides, there is plenty of evidence to say that the doctors are aware of their audience and often address the transcriptionists directly by thanking them and telling jokes.

3. Training Corpora and FP Models

This study used three base and 5 derived corpora:

3.1 Base

- Balanced hand transcribed training corpus (BHT_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_MODEL which controls the process of populating a no-FP corpus with artificial FP's.
- Unbalanced hand transcribed training corpus (UHT_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 and the average frequency with which punctuation is spoken by the doctors.

- Finished transcriptions corpus (NOFP_CORPUS) of 13,537,262 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

- STOCH_FP_CORPUS is a version of the finished transcriptions corpus populated with FP's based on the BIGRAM_FP_MODEL. (FP count: 2, 294, 909)
- RND_FP_CORPUS_3 is derived from the NOFP_CORPUS seeded with FP's with insertion points uniformly distributed in the range between 1 and 6. Here we are hypothesizing that the average length of syntactic phrase is around 3 words. (FP count: 3,867,789)
- RND_FP_CORPUS_5 is derived from the NOFP_CORPUS seeded with FP's with insertion points uniformly distributed in the range between 1 and 10. This corpus represents another estimation of a phrase length interval. (FP count: 2, 707, 842)
- RND_FP_CORPUS_10 is derived from the NOFP_CORPUS seeded with FP's with insertion points uniformly distributed in the range between 1 and 20. This corpus aims at having FP's spaced at sentence length intervals. (FP count: 1,289,796)
- RND_FP_CORPUS_15 is derived from the NOFP_CORPUS seeded with FP's with insertion points uniformly distributed in the range between 1 and 30. This corpus contains a sparse population of FP's with insertion points centered around the average FP frequency derived from UHT_CORPUS. (FP count: 873, 538)
- CBFP_CORPUS is derived from the NOFP_CORPUS seeded with FP's via a method that favors clause boundaries. (FP count: 1, 068, 938)

3.3 FP seeding methods

3.3.1 Stochastic method

Here a bigram model is constructed prior to seeding the NOFP_CORPUS with FP's.

This model contains the distribution of FP's obtained from BHT_CORPUS by using the following formula:

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

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

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

3.3.2 Random method.

This method makes use of Perl's rand() function to determine landing sites for FP's. The frequency of FP's can be regulated by specifying the tails of the uniform distribution produced by the function.

3.3.3 Clause Boundary method

This is a pseudo-random method that makes limited use of linguistic knowledge while populating the NOFP_CORPUS with FP's. Just as the Random method, this one involves flipping a multi-sided coin; however, the landing site for FP depends on finding a clause boundary. For example, if the coin lands on 8 and the eighth word does not designate a clause boundary, the insertion engine will keep searching for a clause boundary from position 8 onwards until it finds one. An FP is inserted and the coin is flipped again.

We used the following lexical items as clause boundary anchors: "period", "that", "which", "if", "whether", "who", "when", "what", "where", "why", "how", "because", "so", "however", "although", "though." We had to take our chances with "that" since it is such a common clause boundary marker, despite the fact that it may also be a demonstrative pronoun.

Using the word "period" as a clause boundary marker presents a problem because the doctors say the word period on average only 20% of the times when the transcriptionists type the dot in the finished transcription. Calculating this number is not straightforward in the absence of corresponding literal and finished transcription corpora. At our disposal we have 500,000 words of literal transcriptions (UHT_CORPUS) which has spoken "periods" but has no punctuation. We also have a 13.5 mil (NOFP_CORPUS) word corpus that has punctuation but does not necessarily represent what was said. The average length of a sentence in the NOFP_CORPUS turns out to be around 10 words. This average length, when applied to the UHT_CORPUS, gives us an estimate of about 50,524 sentences. Given that the word "period" is found in the UHT_CORPUS 10,097 times, we can roughly estimate that about 20% of sentences are terminated with the word "period" actually spoken. This means that to insert an FP at sentence boundaries, the program has to perform the insertion around the "." in NOFP_CORPUS prior to eliminating 80% of punctuation for training the language model.

4. Trigram Language Models

The following trigram models were built using ECRL's Transcriber language modeling tools [6]. Bigram cutoffs were set at 0 and trigram cutoffs were set at 1.

- NOFP_LM was built with the NOFP_CORPUS with no FP's.
- STOCHF_LM ← STOCH_FP_CORPUS.
- RNDFP_LM_3 ← RN_DFP_CORPUS_3
- RNDFP_LM_5 ← RN_DFP_CORPUS_5

- RNDFP_LM_10 ← RN_DFP_CORPUS_10
- RNDFP_LM_15 ← RN_DFP_CORPUS_15
- CBFP_LM ← CBFP_CORPUS.

5. Evaluation Methods

Testing data comes from 23 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.

Recognition accuracy

Recognition accuracy was obtained with ECRL’s HResults tool and is summarized in Table 1.

$$\text{Rec Acc} = \frac{N(\text{hits}) - N(\text{insertions})}{N(\text{total words})}$$

FP Correctness

This metric produces a percentage ratio of correctly recognized FP’s to the total number of FP’s in a given dictation. It is similar to HResults’ (%) Correct measure. The results are summarized in Chart 1.

False FP measure

This measure is useful for identifying real words that got recognized as FP’s. For example, the word “umbilical” may be recognized as “um build”. Chart 2 displays the results.

6. Results and Discussion

Modeling FPs consistently increases total recognition results, even for talkers who use no FP’s. All six FP models we built recorded an increase between 6.083 and 9.15 points. This is in accord with Pakhomov’s [10] report that LMs incorporating FPs systematically decrease the models’ perplexity and increase recognition accuracy.

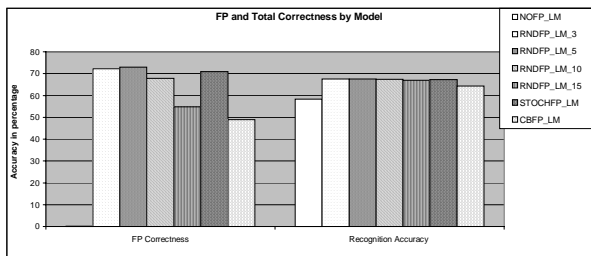


Chart 1: Filled Pause and Total Correctness by Model

Total recognition results for the six FP models do not show significant statistical difference (see Chart 1). However, we introduced an additional measure for the model’s performance – FP correctness (see Chart 1). Measuring correctness of FP recognition separately from overall recognition accuracy allows to evaluate the FP component of the language model. It is important to single out this component because its effects may cancel each other out in the overall recognition accuracy. For example, by increasing FP representation in a language model, one may recognize some of the FP’s which went unrecognized before; however, some of the other words that were recognized

correctly before may begin to be misrecognized as FP’s. The overall recognition accuracy in a situation like this may not change which would make it impossible to detect a change in performance. Measuring FP accuracy on the six models shows a considerable difference among some of the models’ performance. FP correctness appears to grow proportionately with the frequency of FP’s in the training corpus. The top four FP models - RNDMFP_LM_3, RNDMFP_LM_5, RNDMFP_LM_10 and STOCHF_LM - do not exhibit a significant variation between them, but as a group appear to perform better than RNDFP_LM_15 and CBFP_LM. False FP measure, however, shows that RNDFP_LM_3 and RNDFP_LM_5 produce significantly more false FP’s than any other model, which leaves RNDFP_LM_10 and STOCHF_LM as the two optimal models in terms of the trade-off between overall recognition correctness, FP correctness and false FP rate.

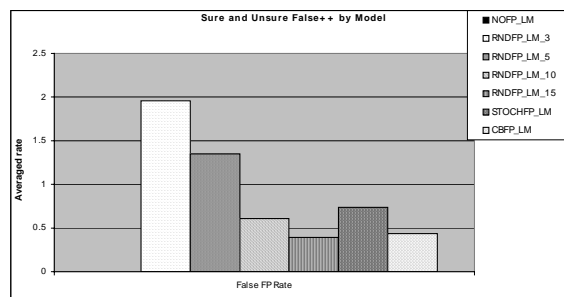


Chart 2: Sure/Unsure False Positive Rate by Model

The frequency of FP’s represented in RNDFP_LM_10 corresponds roughly to FP’s inserted around sentence boundaries. Our other model, CBFP_LM, intended to reflect linguistic knowledge about clause boundaries showed only satisfactory results. Nevertheless, the fact that RNDFP_LM_10 (trained on a corpus with FP’s uniformly distributed around average sentence length) is among the most optimal models suggests that the clause boundary approach is not without promise.

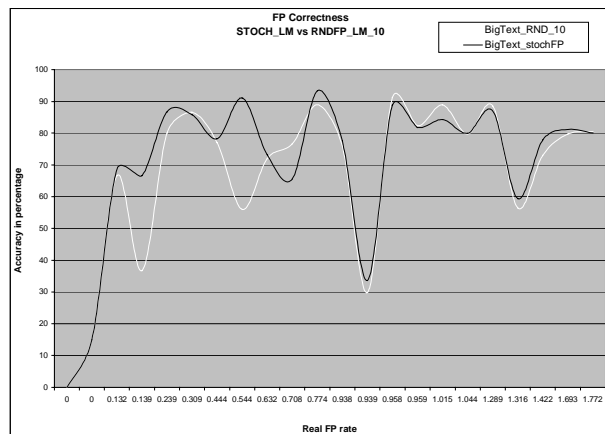


Chart 3: FP Correctness – STOCHF_LM vs RNDFP_LM_10 sorted left to right by FP Rate

We correlated [the results](#) for the top two models, STOCHFPM and RNDFPM₁₀, in terms of total recognition accuracy, FP correctness and false FP rate to actual FP rate (see Chart 3). STOCHFPM (with 2,294,909 FP's) performs slightly better, although not significantly, on low FP users (FP rate smaller than 10%). For high FP users, RNDFPM₁₀ seemed to yield slightly better results. That model has 1,289,796 FP's. For these two models, the raw number of FP's in them is not a good predictor for group performance (low vs high FP users). Thus, we are concluding that it is not the raw FP frequency but rather a combination of frequency and pattern of distribution of the FP's in the training corpus that correlates with FP recognition correctness. An intuitive suggestion to populate a corpus with fewer FP's when tailoring a model to low FP users would not necessarily [yield better results](#). [On the contrary, it might even hurt the FP correctness rate](#).

STOCHFPM implicitly incorporates linguistic knowledge since it is based on a bigram model built from a hand-transcribed corpus that includes FP's. However, the linguistic knowledge is not generalized in a scheme that we could explicitly use. Our 'true' linguistic model, CBFP_{LM}, obviously does not incorporate an optimal amount of linguistic knowledge. Despite its mediocre performance, we do think that a balanced patterned representation of FP's in the training corpus will eventually yield better results than the purely random FP insertions.

Another interesting observation has to do with the fact that recognition accuracy improves with an FP model compared to NOFP model for talkers who use no FP's. This side effect may be linked to fact that introducing FP's into the training corpus decreases the model's perplexity [9] and results in better recognition overall, FP's or no FP's. We are planning to investigate this issue further.

7. Conclusion

The current results of our study indicate that for speech recognition purposes FP distribution can be modeled stochastically or with a uniform distribution centered around average sentence length. We were unable to obtain satisfactory results by seeding the training corpus with FP's based on limited linguistic knowledge, which we attribute to under representation of clause boundary rules. To improve the representation one would have to make use of a parser. We have also shown that using recognition accuracy as the only measurement to determine the goodness of an FP model is not sufficient – other methods such as FP correctness and false FP detection must be used for accurate evaluation.

ACKNOWLEDGEMENTS

Our thanks go to Joan Bachenko, PhD, and Michael Moon, PhD, Linguistic Technologies, Inc., Edina, Minnesota for their invaluable advice and warm encouragement.

Note: the authors can be reached at pakh0002@tc.umn.edu and savo0014@tc.umn.edu

REFERENCES

- [1] Chen, S., Beeferman, Rosenfeld, R. 1998. "Evaluation metrics for language models," In DARPA Broadcast News Transcription and Understanding Workshop.
- [2] Christenfeld, N, Schachter, S and Bilous, F. 1991. "Filled Pauses and Gestures: It's not coincidence," Journal of Psycholinguistic Research, Vol. 20(1).
- [3] Cook, M. 1977. "The incidence of filled pauses in relation to part of speech," Language and Speech, Vol. 14, pp.135-139.
- [4] Cook, M. and Lalljee, M. 1970. "The interpretation of pauses by the listener," Brit. J. Soc. Clin. Psy. Vol. 9, pp. 375-376.
- [5] Cook, M., Smith, J, and Lalljee, M 1977. "Filled pauses and syntactic complexity," Language and Speech, Vol. 17, pp.11-16.
- [6] Valtchev, V. Kershaw, D. and Odell, J. 1998. The truetalk transcriber book. Entropic Cambridge Research Laboratory, Cambridge, England.
- [7] Lalljee, M and Cook, M. 1974. "Filled pauses and floor holding: The final test?" Semiotica, Vol. 12, pp.219-225.
- [8] Maclay, H, and Osgood, C.1959. "Hesitation phenomena in spontaneous speech," Word, Vol.15, pp.19-44.
- [9] Pakhomov, S. 1999. "Modeling Filled Pauses in Medical Dictations." Proc. ACL'99.
- [10] Shriberg, E. E. 1994. Preliminaries to a theory of speech disfluencies. Ph.D. thesis, University of California at Berkeley.
- [11] Shriberg, E.E. and Stolcke, A. 1996. "Word predictability after hesitations: A corpus-based study," In Proc. ICSLP.
- [12] Shriberg, E.E. 1996. "Disfluencies in Switchboard," In Proc. ICSLP.
- [13] Shriberg, E.E. and Bates, R. and Stolcke, A. 1997. "A prosody-only decision-tree model for disfluency detection" In Proc. EUROSPEECH.
- [14] Siu, M. and Ostendorf, M. "Modeling disfluencies in conversational speech," Proc. ICSLP, 1996.
- [15] Stolcke, A and Shriberg, E.1996. "Statistical language modeling for speech disfluencies," In Proc. ICASSP.
- [16] Swerts, M, Wichmann, A and Beun, R. 1996. "Filled pauses as markers of discourse structure," Proc. ICSLP.