Adding Robustness to Language Models for Spontaneous Speech Recognition

Jacques Duchateau, Tom Laureys, Patrick Wambacq

Katholieke Universiteit Leuven - ESAT
Kasteelpark Arenberg 10
B-3001 Leuven, Belgium
E-mail: Jacques.Duchateau@esat.kuleuven.ac.be

Abstract

Compared to dictation systems, recognition systems for spontaneous speech still perform rather poorly. An important weakness in these systems is the statistical language model, mainly due to the lack of large amounts of stylistically matching training data and to the occurrence of disfluencies in the recognition input. In this paper we investigate a method for improving the robustness of a spontaneous language model by flexible manipulation of the prediction context when disfluencies occur. In the case of repetitions, we obtained significantly better recognition results on a benchmark Switchboard test set.

1. Introduction

The automatic recognition of spontaneous speech is currently one of the main topics in speech research. Practical applications include voice operated telephone services, automatic closed captioning for TV programs, automatic transcription of meetings, etc. Yet, the recognition accuracy of freely spoken language is quite poor when compared to that of dictated speech: while the word error rate (WER) for large vocabulary speaker-independent dictation is about 5%, the WER for spontaneous speech recognition ranges from 15% for broadcast news [1, 2] to 40% for meeting and telephone conversation transcription [3].

One of the main reasons for this discrepancy is the lack of a sufficient amount of stylistically matching training data to estimate spontaneous language models. Written transcripts of casual language use are rather scarce, while typical large vocabulary stochastic language models rely on vast amounts of training material [4]. The occurrence of disfluencies in casual speech makes a spontaneous language model even less robust. This paper focuses on the latter problem.

In the literature different approaches to spontaneous language modeling have already been pursued. [5] tried to incorporate knowledge of discourse theory: sentences typically start with given information whereas new information comes at the end. Correspondingly, two *expert* language models were trained on the relevant sentence parts, yielding a slight 0.3% absolute improvement in WER for recognition of spontaneous telephone conversations (Switchboard). Disfluencies almost always occurred in the sentence's given information part. [6] explore N-best list rescoring on the basis of chunking information. The underlying motivation is that the coverage of the chunker bears information in order to discriminate between syntactically acceptable and syntactically anomalous recognition hypotheses. The technique reduced the WER by 0.3% absolute on Switchboard. Finally, [7] report on dealing with disfluencies in language modeling by editing the prediction context. More specifically, the prediction context for a newly hypothesized word is *cleaned up* by removing the disfluencies in it. The improvement in WER on Switchboard is, parallel to the other approaches, not really significant. The research described in this paper extends the latter work by implementing a more flexible manipulation of the prediction context: disfluencies are only removed from the context when they do not contain informational value.

The paper is organized as follows. First, we discuss the investigated disfluencies and the proposed model to handle them. Next, the experimental set-up is described and results on the Switchboard task are given. Finally, we conclude and discuss future research on the topic.

2. Handling disfluencies

2.1. The investigated disfluencies

As mentioned above, one of the features that distinguishes spontaneous from read speech is the occurrence of disfluencies. The disfluency types we focus on in this work are listed below:

repetitions: *That is what **what** I think.*

hesitations: *That is what **um** I think.*

sentence restarts: *That is what um... Yeah I think so.*

About 85% of the disfluencies in our train and test corpus (Switchboard, cf. infra) are of the three types listed above [8]. Therefore we suppose that the behavior of our model for disfluencies will be reflected adequately by the selected disfluency types.

One of the hypotheses explaining the difficulty of spontaneous language modeling by means of N-grams

points explicitly to disfluencies: as N-grams base their word prediction on a local context of N-1 previous words, intervening disfluencies render this context less uniform. Or put differently, the prediction of a next word would be more accurate if based on a context from which disfluencies are removed and which is extended to the left with regular words to make up for the removed disfluencies. So when using a trigram language model (LM) in the case of the hesitation mentioned above, we hypothesize that *I* would be better predicted by the context *is what* than by *what um*. The disfluencies themselves are predicted in the same way as regular words.

Yet, as shown by [9] and [10], in some cases disfluencies *are* good predictors for following words. Hesitations, for example, sometimes tend to precede less frequently used words (depending, among other factors, on the position of the hesitation in the sentence). In addition, repetitions are not always grammatically incorrect (e.g. *I hope that that work is done.*). So simply removing disfluencies from the prediction context seems too crude. In our model we tried to incorporate this observation by allowing the system to pick the most probable option when both a context with disfluency and a cleaned-up context are available.

2.2. The proposed model

We explain in detail how the proposed model works by taking the case for repetitions as an example. The model for repetitions is sketched in figure 1. As can be seen on the figure, we assume that a trigram language model is used. The upper path illustrates the normal LM procedure. Suppose that word *B* is repeated, then the prediction of the next word *C* is based on the context *B B*. The removal of the repetition is demonstrated by the lower path. The prediction of *C* is made on the basis of the modified context *A B*; the repeated word *B* is removed.

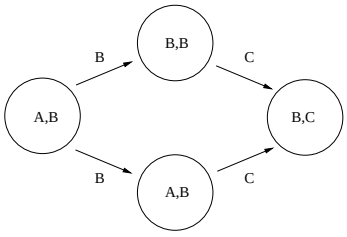

Figure 1: The model for repetitions

In section 3, we will compare the reference system, which always follows the upper path, with a system that always removes the disfluency from the context (thus always follows the lower path). This comparison was also conducted in [7]. Additionally, we also investigate a system that selects the most probable prediction context. In that case the prediction of *C* is based on the most probable of both contexts mentioned, and depending on the situation the upper path or the lower path is chosen.

The analogous models for hesitations (symbol *uh*) and sentence restarts (context <*S*>) are depicted in figures 2 and 3 respectively. The figure shows that in these cases, it takes one word more for both options to join again. It should be noted that in the model for sentence restarts, a sentence restart is only allowed following a hesitation although in spontaneous speech a sentence can restart at any point. However, a pilot experiment described in [11] showed that a restart following any word overgenerates hypotheses and worsens recognition.

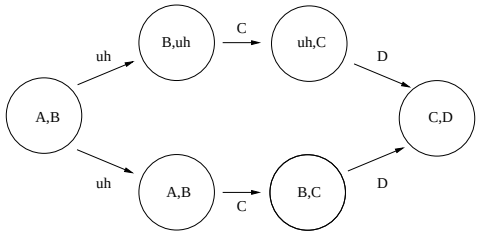

Figure 2: The model for hesitations

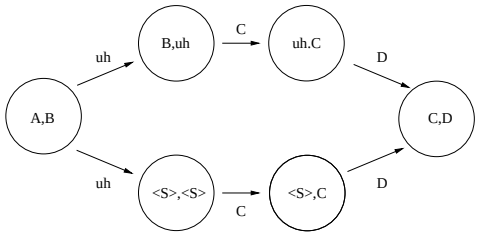

Figure 3: The model for sentence restarts

Before turning to the real recognition experiments, we set up a small-scale experiment to investigate whether the probabilities in the trigram language model, estimated on a rather small word text database (only 3M words, cf. infra), were reliable enough to distinguish between the different prediction contexts compared in the proposed models. We did this by analyzing sentence restarts after the hesitation *uh* in a Switchboard test set. The test set contained 72 occurrences of *uh* in the middle of the sentence. From a manual examination we learned that in 18% of these cases the sentence restarted following the hesitation, and in the remaining 82% of the cases the sentence just went on.

Next, we made the language model choose between both contexts with the model depicted in figure 4. We found that both for the sentence restarts and for the continued sentences, the LM was able to select the correct transition in the model in 84% of the cases. As this classification task result is based purely on the LM (not on an optimized classifier), it clearly indicates that most information on the optimal LM prediction context can be found in the trigram language model.

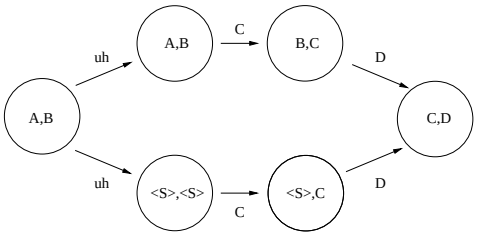

Figure 4: The model for the language model evaluation

3. Experiments and results

3.1. Experimental setup

The proposed models for handling disfluencies were evaluated by means of recognition experiments with the ESAT speech recognizer and based on the Switchboard corpus, a collection of informal telephone conversations in American English [12]. This section describes the baseline recognition system and the benchmark Switchboard test set.

Gender independent *acoustic models* were estimated on the 310 hours of Switchboard-1 data. A global phonetic decision tree defines 8K tied states in the cross-word context dependent and position dependent models. Each tied state is modeled with a mixture of on average 220 tied gaussian distributions from a total set of 117K different gaussians.

A Good-Turing smoothed trigram *language model* was built on the basis of the 3M words in the Switchboard-1 conversation transcripts. The recognition lexicon consisted of the 27K words in the Switchboard-1 training data.

For the test set, the 2001 HUB5 benchmark was used, more particularly the part that corresponds to Switchboard-1 (data that was of course excluded from the training set). This test set consists of 20 phone calls: in total almost 2 hours of data, or 1718 sentences with 20K words. Reference transcriptions and scoring software for this benchmark can be found at *ftp://jaguar.ncsl.nist.gov/lvcsr/mar2001*.

The *decoder* of the recognition system is based on a single pass time synchronous beam search algorithm (no speaker adaptation was used). The baseline recognition result on the 2001 HUB5 test set is 29.8% WER. The recognizer then runs 4 times slower than real time (on a 2.8 GHz Pentium 4 processor). Real time recognition (using smaller acoustic models and investigating fewer hypotheses in the search) results in a WER of 31.6%.

For the experiments in [11], the models for handling disfluencies were implemented directly into the single pass recognition. For the experiments in this paper however a more flexible (flexible concerning the planned incorporation of acoustic-prosodic information) two pass strategy was adopted, generating graphs with hypotheses in the first pass and rescoring them in the second.

3.2. Results and discussion

The resulting WERs for the recognition experiments are summarized in table 1. For each of the three disfluency types, three types of context manipulation were investigated: leaving the context unchanged (the baseline experiment), changing the context according to the model, and choosing the most probable of the two former options.

	unchanged	changed	choice
repetition	29.8%	29.7%	29.6%
hesitation	29.8%	29.9%	29.8%
restart	29.8%	29.8%	29.9%

Table 1: WERs for the different disfluency types with varying context manipulation: full test set

Given the size of the test set, the differences between the WERs seem to be insignificant at first sight. But it is not really fair to say so as only about 5% of the words in the test set are disfluencies. In practice the investigated models influence only about 20% of the sentences.

In order to find an appropriate subset of sentences to do the evaluation, it's not a good idea to evaluate the investigated models on the sentences which contain disfluencies only as the models can introduce errors in other sentences. However this problem can be solved by evaluating only the sentences for which there is a difference between the recognition result without and with the model. Doing so, table 2 is found. Note that the result with unchanged context varies as the selected test set depends on the experiment. The fact that the results are worse then on the full test set is probably due to the selection of typically long sentences in the test set.

	unch'd	changed	unch'd	choice
repetition	35.2%	34.8%	**36.7%**	**35.1%**
hesitation	33.2%	33.9%	36.8%	37.4%
restart	34.9%	35.4%	35.7%	36.5%

Table 2: WERs for the different disfluency types with varying context manipulation: partial test set

In table 2 the results are clearer. On the left, the results with changing context are given, on the right those with choice in context. Although still only one result is really significant (the improvement using the model for repetitions with choice in context), we can conclude that the proposed model improves the robustness of the system in case of repetitions, but slightly deteriorates results for hesitations and restarts.

This different behavior is probably due to the detection of the disfluency. For hesitations and restarts this detection is weak: it is simply based on the recognition of the *uh* word, which is modeled in the current system as a choice between 4 short phonetic strings. Using this acoustic model, a hesitation can be hypothesized easily. So this generates many possible prediction context changes. This problem can probably be solved by improving the hesitation detection, for instance by using acoustic-prosodic cues that point to the presence of a disfluency.

4. Conclusions and future research

In this paper we investigated whether spontaneous language modeling can benefit from a specific approach to disfluencies. We tried to improve on the robustness of a plain trigram LM by manipulating prediction contexts containing repetitions, hesitations or restarts.

In case of repetitions, we found that the recognition can be improved significantly by offering the recognition system the choice between removing or not removing the disfluency from the prediction context. However for hesitations and restarts this method results in a small deterioration of the recognition rate.

In a first step in our future research we will try to solve this problem by improving the detection of hesitations using additional acoustic-prosodic information.

Further, we will set up experiments on the inclusion of additional LM training material. Perplexity measures can indicate which parts of written text data are grammatically or stylistically close to spontaneous speech. Adding those texts when training the LM can improve the statistics when disfluencies are removed from the prediction context and at the same time lead to a more accurate automatic context selection.

5. Acknowledgments

This research was supported by the ADV/STWW/000151 ATraNoS project (http://atranos.esat.kuleuven.ac.be) and by the IST-2001-38299 MUSA project.

6. References

[1] P. Beyerlein, X. Aubert, R. Haeb-Umbach, M. Harris, D. Klakow, A. Wendemuth, S. Molau, M. Pitz, and A. Sixtus, "The Philips/RWTH system for transcription of broadcast news," in *Proc. European Conference on Speech Communication and Technology*, vol. II, Budapest, Hungary, Sept. 1999, pp. 647–650.

[2] J. Gauvain, L. Lamel, G. Adda, and M. Jardino, "Recent advances in transcribing television and radio broadcasts," in *Proc. European Conference on Speech Communication and Technology*, vol. II, Budapest, Hungary, Sept. 1999, pp. 655–658.

[3] H. Yu, T. Tomokiyo, Z. Wang, and A. Waibel, "New developments in automatic meeting transcription," in *Proc. International Conference on Spoken Language Processing*, vol. IV, Beijing, China, Sept. 2000, pp. 310–313.

[4] G. Adda, M. Jardino, and J. Gauvain, "Language modeling for broadcast news transcription," in *Proc. European Conference on Speech Communication and Technology*, vol. IV, Budapest, Hungary, Sept. 1999, pp. 1759–1762.

[5] K. Ma, G. Zavaliagkos, and M. Meteer, "Bi-modal sentence structure for language modeling," *Speech Communication*, vol. 31, no. 1, pp. 51–67, 2000.

[6] K. Zechner and A. Waibel, "Using chunk based partial parsing of spontaneous speech in unrestricted domains for reducing word error rate in speech recognition," in *Proc. 17th Conference on Computational Linguistics (COLING/ACL'98)*, Montreal, Canada, Aug. 1998, pp. 1453–1459.

[7] A. Stolcke and E. Shriberg, "Statistical language modeling for speech disfluencies," in *Proc. International Conference on Acoustics, Speech and Signal Processing*, vol. I, Atlanta, U.S.A., May 1996, pp. 405–408.

[8] E. Shriberg, "Disfluencies in Switchboard," in *Proc. International Conference on Spoken Language Processing*, vol. Addendum, Philadelphia, U.S.A., Oct. 1996, pp. 11–14.

[9] M. Siu and M. Ostendorf, "Modeling disfluencies in conversational speech," in *Proc. International Conference on Spoken Language Processing*, vol. I, Atlanta, U.S.A., Oct. 1996, pp. 386–389.

[10] E. Shriberg and A. Stolcke, "Word predictability after hesitations: a corpus-based study," in *Proc. International Conference on Spoken Language Processing*, vol. III, Philadelphia, U.S.A., Oct. 1996, pp. 1868–1871.

[11] J. Duchateau, T. Laureys, K. Demuynck, and P. Wambacq, "Handling disfluencies in spontaneous language models," in *Computational Linguistics in the Netherlands*, ser. Language and Computers. Studies in Practical Linguistics, T. Gaustad, Ed. Amsterdam (The Netherlands) and New York (U.S.A): Rodopi, 2003, pp. 39–50.

[12] J. Godfrey, E. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *Proc. International Conference on Acoustics, Speech and Signal Processing*, vol. I, San Francisco, U.S.A., Mar. 1992, pp. 517–520.