MODELING DISFLUENCY AND BACKGROUND EVENTS IN ASR FOR A NATURAL LANGUAGE UNDERSTANDING TASK

R. C. Rose and G. Riccardi
AT&T Labs - Research, 180 Park Ave., Florham Park, NJ 07932
(rose,dsp3)@research.att.com

ABSTRACT

This paper investigates techniques for minimizing the impact of non–speech events on the performance of large vocabulary continuous speech recognition (LVCSR) systems. An experimental study is presented that investigates whether the careful manual labeling of disfluency and background events in conversational speech can be used to provide an additional level of supervision in training HMM acoustic models and statistical language models. First, techniques are investigated for incorporating explicitly labeled disfluency and background events directly into the acoustic HMM. Second, phrase–based statistical language models are trained from utterance transcriptions which include labeled instances of these events. Finally, it is shown that significant word accuracy and run–time performance improvements are obtained for both sets of techniques on a telephone–based spoken language understanding task.

1 INTRODUCTION

It is well known that the existence of non–speech events in spontaneous speech utterances has some effect on automatic speech recognition (ASR) performance. The work described in this paper attempts to deal with this problem by training both acoustic HMMs and phrase–based n–gram language models using transcriptions of utterances that include labeled occurrences of these events. The acoustic models and language models trained in this manner result in an LVCSR system that provides a significant advantage over existing ASR systems. This advantage has been gained partly from the manual labeling of these events and partly from modeling techniques which exploit the localized effects of these events on the surrounding utterance. In most existing ASR systems, the training of the background event acoustic models is done in an unsupervised mode. Furthermore, there is no attempt to obtain a probabilistic characterization of how these events occur in the context of word sequences. The final result of the paper demonstrates the importance of more sophisticated models of non–speech events by comparing the ASR performance of systems that do explicitly model these events with systems that do not.

There are two types of manually labeled non–speech events that are addressed in this paper. "Disfluency events" include instances of filled pauses, word fragments, and hesitations which are generally associated with disfluent speech. "Background events" refer to instances of human generated non–speech noise including breath noise, lip-smacks, and laughter as well as background noise and background speech. Other background events include echoed prompts that may overlap the user's utterance. It is not immediately obvious that using these events will result in better automatic speech recognition (ASR) performance. There are issues relating to statistical robustness, whether a given spontaneous speech or background event occurs often enough to train the parameters of a unit to represent the event. There are also issues of acoustic robustness, whether there is a stable and consistent acoustic realization of the event. Finally, there are issues relating to the consistency and quality of the labeling procedures themselves.

It is shown that supervised training of acoustic models representing fifteen classes of disfluency and background events can outperform unsupervised training of general background HMMs. Section 4 describes the experimental study associated with supervised training of models for disfluency and background events. The results of this study show that significant improvement in word accuracy (WAC) can be obtained simply by including these units in an optional "between-word" loop within the recognition network as illustrated in Figure 1a.

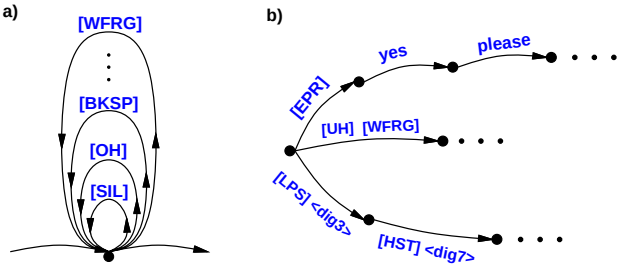

Figure 1: *a)* **Inclusion of all labeled background events (LBEs) in a single "between–word" loop.** *b)* **Portion of phrase–based LM trained from LBE annotated text.**

The effect of including these labeled events in training stochastic language models is also investigated. First, the inclusion of these events in the calculation of n–gram probabilities is explored. Second, these events are incorporated into phrase–based n–gram models as is illustrated in Figure 1b. It is shown that significant performance improvements can be gained through the formation of phrases that combine disfluency or background events with vocabulary words.

There are several previously published studies analyzing the effects of incorporating additional labeled information concerning pronunciation variation and disfluencies in configuring LVCSR systems for spontaneous speech tasks. Stolcke and Shriberg estimated language model probabilities for words occurring after various types of disfluencies by conditioning the words on the intended fluent text instead of the disfluent event [5]. Our study differs from this previous work in that it is more concerned with analyzing the effects of *increasing* the degree to which language model probabilities are conditioned on explicit representations of spontaneous speech and background events. Another study

analyzed the effects of using labeled occurrences of filled–pauses and other non–speech events for training HMMs and n–gram language models for the Broadcast News task [2]. Our study performs a similar analysis using a wider class of labeled events and a broader set of performance metrics on a natural language understanding task.

The outline of the paper is given as follows. Section 2 briefly describes the natural language understanding task and how non–speech events are represented in utterance transcriptions. Issues relating to incorporating LBEs into a phrase–based statistical language model are discussed in Section 3. The effect of labeled background events on the acoustic model training and the results of an experimental study evaluating the effects of LBEs on ASR performance are given in Section 4.

2 SPOKEN DIALOG TASK

The speech corpus that is used in this study is described in two parts. The first part briefly describes the telephone based natural language understanding task. This includes a summary of the speech corpus that was created from a subset of the responses to particular queries in the task. The second part of the section enumerates the manually labeled disfluency and background events which are referred to below as "labeled background events" (LBEs).

2.1 Natural Language Understanding Task

The task considered in this work involves a spoken dialog that is carried out with a user to interpret the user's telephone call according to a set of call–types and to obtain from the user all necessary information associated with the given call type [1]. The call–types themselves correspond to a set of actions relating to the routing of the incoming call. Examples of these call–types include *collect*, *credit card*, and *third party billing*, with an additional "*other*" type to handle calls that do not correspond to those that have been defined. The queries that are presented to the user range from being very unconstrained to being highly focused. The techniques in this paper are applied to utterances that correspond to responses to two different queries. The first set of utterances are responses to the open-ended query "How may I help you?", and will be referred to as greeting utterances [3]. The second set of utterances were made in response to the prompt "May I have your card number please?", and will be referred to as card number utterances.

The training data used in the experimental study was obtained from two separate speech corpora that were collected at separate times. They are referred to collectively as the "How May I Help You?" (HMIHY) corpus. The first consists of human–human interactions between customers and a human operator [3]. The second data set consists of human–machine interactions between customers and an automated system [1]. There were a total of 10600 speech utterances used for training subword acoustic HMMs and 16159 utterance transcriptions for training stochastic language models.

2.2 Representation of Labeled Events

Table 3 enumerates the labeled background events (LBEs) in the HMIHY corpus. The first column of Table 3 contains the total number of occurrences of each LBE class in the speech utterances. The advantage of incorporating these events is highly dependent on the quality of the process by which human labelers annotate the transcriptions. This quality is maintained through the use of a common speech labeling guide, the use of a set custom designed set of software tools for speech labeling, and the practice of performing two labeling passes over each set of utterances.

While there are dozens of possible label codes that can be used to annotate the transcriptions, these are compressed into a set of fifteen symbols for this study. Disfluency events include filled pauses, word fragments, and hesitations. Separate symbols are included for each of six possible filled pauses ([ah], [eh], [uh], [um], [oh], [er]). There are over 1200 occurrences of word fragments in the transcriptions. However, because of the large acoustic variability of word fragments and the high level of difficulty in obtaining consistent labeling for these events, it was initially decided to replace word fragment occurrences by a single word fragment symbol [wfrag]. As a result, the word fragments *tr-* in "I'm tr- trying to make a call ..." and *connec-* in "yeah can you conne- connect me with information ..." were both replaced by [wfrag]. A separate symbol, [hst], is also included for hesitations which occur most often in the form of silence intervals greater than one second long in the context of disfluencies.

Other labeled events include both human generated and background noises. Human generated non-speech noises like laughter [lgh], lip–smack [lps], and breath [brth] are given unique symbols. Additional background related noises were collectively categorized as "background speech noise" ([bksp]) and "non–speech noise" ([nspn]). Finally, symbols for operators' speech ([oper]) in the human–human interactions and echoed prompt ([epr]) in the human–machine interactions were included.

Summary of Background Events in HMIHY Corpus		
Background Event	Total Counts	K-L Distance
Filled Pauses (6)	7189	1.72
Word Fragments	1265	1.73
Hesitations	792	1.97
Laughter	163	1.16
Lipsmack	2171	1.47
Breath	8048	2.33
Non-Speech Noise	8834	1.24
Background Speech	3585	1.46
Operator Utt.	5112	1.20
Echoed Prompt	5353	1.78

Table 1: Counts of the number of occurrences of labeled background events (LBEs) in training utterances, and a K-L distance which represents the power of the LBEs as predictors of words in an utterance.

3 LABEL–CLASS DEPENDENT LANGUAGE MODELS

Language models for speech recognition are generally defined over a vocabulary of words and do not explicitly characterize the occurrence of the non–speech events that were described in Section 2. The parameters for these models are trained from a corpus $\mathcal{T}$ containing utterance transcriptions $W = w_1, \ldots, w_N$, $w_i \in V$, where V is the dictionary of lexical units (words) drawn from $\mathcal{T}$. This section investigates the potential improvements that might be realized by incorporating the non-speech information represented by the LBEs into stochastic language modeling techniques for ASR. In this section we analyze the distribution of the LBEs in the context of transcribed word sequences. It is shown that these events have a non–uniform, predictable pattern of occurrence in speech utterances. It is also shown that LBEs often serve as *phrase markers*.

3.1 LBEs and language mutual information

The experimental analysis performed in this section has two goals. The first is to determine whether the occurrence of transcribed non-speech events can serve as reliable predictors of words in an utterance. This is approached from an information theoretic point of view by measuring whether conditioning the occurrence of words on the occurrence of LBEs can reduce the language entropy. The second goal

is to determine whether occurrences of these non-speech events are predictable from word contexts in the utterance transcriptions. It is important that both of these goals be satisfied if any significant gains are to be achieved from a language model defined over both words and LBEs.

Mutual information is used to determine if LBE conditioned word occurrence probability reduces the language entropy, $H(W)$, for a word sequence W. Let us denote the LBEs as $\delta_i \in \Delta$, where Δ is the dictionary LBEs. Then the mutual information between a word sequence W and a sequence of words conditioned on LBE context δ is given by

$$I(W, \delta) = H(W) - H(W|\delta) \quad (1)$$

where $H(W|\delta)$ is the LBE conditioned entropy of W.

A slightly different measure is used to determine whether LBE occurrences can be predicted from their preceding lexical contexts. Let us define a generic sequence that is augmented to include both words and LBEs by $\hat{W} = d_1, \ldots, d_N$, with $d_i \in V \cup \Delta$. For this augmented sequence, we can compute the conditional entropy $H(\hat{W}|\delta)$. Then we can define a new quantity

$$G(W, \delta) = H(W) - H(\hat{W}|\delta) \quad (2)$$

as the reduction in entropy that is achieved both by augmenting W to obtain $\hat{W}$ and also conditioning the probability of d_i on the LBEs. The quantity $G(W, \delta)$ in Equation 2 is not a mutual information measure, and can in fact take on negative values. It is interesting to note that $I(W, \delta) \geq G(W, \delta)$, and that information is only gained by defining a language model over the augmented lexicon if $G(W, \delta) \geq 0$.

The quantity $G(W, \delta)$ was computed over the entire set of transcriptions in order to obtain a lower bound on the mutual information for two types of language models. The first was a word–based bigram model which *always* exploits one symbol, v_i, to predict the next v_{i+1}. The second was a phrase–based bigram model which provides a variable time window (typical range is 1-20) corresponding to automatically acquired phrases [4, 3]. Table 2 displays the estimated values of the quantities in Equation 2 for the word–based and phrase-based language models. There are two important points that can be made from Table 2. The first is that the estimated entropy $H(W)$, and consequently the perplexity, of the phrase–based bigram is always lower than the word–based bigram. This has been well documented for this task [3]. Therefore, all further experiments described in Section 4 will make use of phrase–based language models. The second point that can be made from Table 2 is that $G(W, \delta)$ is significantly greater than zero in both cases which suggests that information is indeed gained by augmenting the original lexicon with the LBE units. This also suggests that the LBEs may be acting as *phrase markers* which supports psycho–linguistic evidence obtained elsewhere [5].

	Word Bigram	Phrase Bigram
$H(W)$	4.8	4.2
$H(\hat{W}\|\delta)$	4.5	4.0
$G(W, \delta)$	0.3	0.2

Table 2: Entropies and Mutual Information Lower Bounds for word and phrase bigrams.

3.2 LBE probability distributions

A simple analysis of the LBE annotated transcriptions was performed to illustrate the information lost when these events are not incorporated in the language model as is illustrated by the "between–word" loop in Figure 1. It is easy to cite specific examples of utterances like telephone or credit card numbers where filled pause or breath noise events do *not* occur randomly over the entire utterance. Similarly, it is well known that word fragments tend to occur near the beginning of a phrase fragment. In order to quantify the correlation between LBE probability distribution and word contexts, the Kullback-Leibler distance between the conditional distribution of word contexts given a particular LBE symbol and the uniform word distribution was computed. The K-L distance is simply the difference between the conditional entropy, $H(d|\delta_i) = -\sum_j p(d_j|\delta_i) \log p(d_j|\delta_i)$, and its upper bound, $logM_i$, where M_i is the number of distinct symbols d_j following δ_i. The second column of Table 3 displays this K-L distance for a range of LBE classes. Over a vocabulary size of almost $3.6K$ words, the most frequent labeled event, "breath", precedes at most 274 symbols d_i. As a result, the K-L distance for "breath" as displayed in Table 3 is fairly high. This, supports the hypothesis that LBEs are *not* randomly occurring acoustic events, and suggests that acoustic and language model training should take into account their acoustic and word contexts.

4 LABEL–CLASS DEPENDENT LVCSR

This section discusses the effect on ASR performance of LBE based acoustic and language models. The training of acoustic and language models will be briefly described along with the initial configuration of the speech recognition system. Finally, experimental results will be presented on two different classes of utterances.

4.1 Baseline System

Both the acoustic and language models in the baseline ASR system were trained using utterance transcriptions where the labeled background events were removed. As a result, the baseline acoustic HMM did not include LBE units. Instead two silence HMM units were trained by including them as optional units in the network during forward–backward training. This is a common scenario where units representing acoustic background information are trained in an "unsupervised" mode. There were a total of 10600 utterances used for acoustic model training which corresponded to approximately twelve hours of speech. Context–independent subword HMMs were used in the system with three states per model and 32 mixture components per state. In addition, dedicated models were trained for the digits zero through nine using eight to ten states per digit.

Phrase-based language models have been found to improve over word–based n-grams by considering a variable length time window spanning over long lexical contexts. The process that learns how to vary the window length is automatic and the size of the stochastic model is comparable to the corresponding word–based n-gram [4, 3]. A phrase based stochastic bigram language model was trained from a corpus of 16159 utterance transcriptions containing 18% natural number tokens. Each instance of a natural number was replaced with a non–terminal symbol to speed up convergence of the phrase acquisition algorithm. LBEs were not included in the transcriptions used for training the baseline language model, and the resulting model had a perplexity of 18.4. The ASR word accuracy (WAC) for this baseline system is given in the first row of Table 3 for both the greeting and card number utterances. Note that for each of the three system configurations in Table 3, the same language and acoustic model is always used for both classes of utterances. There were a total of 762 test utterances for the greeting data set and 342 utterances in the card number data set. A WAC of 58.7 percent was obtained for the greeting utterances and 87.7 percent was obtained for the card number data set.

ASR Word Accuracy			
System Configuration		Test Corpora	
HMM	Language Model	Greeting	Card Number
Baseline	Baseline	58.7	87.7
LBE	Baseline	60.8	88.1
LBE	LBE	60.8	89.8

Table 3: Table demonstrating the effects of using the labeled background events (LBEs) for training HMMs and language models.

4.2 HMM Acoustic Modeling of LBEs

Dedicated acoustic models were trained for each of the fifteen LBEs described in Section 2. The utterance transcriptions used in training were processed so that all non–speech events were compressed into the fifteen LBEs. The topology of LBE models was determined empirically. Anywhere from three to six HMM states were assigned to a LBE unit depending on the average duration of the unit and the number of occurrences of the unit in the training data. The parameterization of the subword unit models and digit models was identical to that used in the baseline system. It is clear from the counts in Table 3 that there are a sufficient number of occurrences of all LBEs to provide reasonably robust estimates of LBE based HMM parameters.

In order to isolate the effects of acoustic modeling of LBEs from the effects of including LBEs in the language model, the language model in Figure 1a was used. The loop in Figure 1a simply indicates that all fifteen LBEs were included in a single "between–word" loop for the baseline phrase–bigram language model described in Section 4.1. There was no retraining of the language model probabilities. The performance of this system using the LBE based acoustic model and baseline language model is given in the second row of Table 3. While there is a significant performance increase for both test sets, the increase in word accuracy is much larger for the less constrained utterances in the greeting test set. For this case, Table 3 shows a performance improvement of 2.1 absolute percentage points.

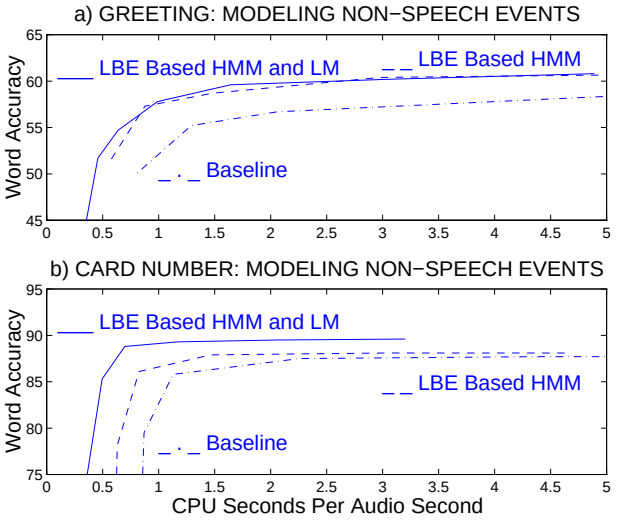

Figure 2: Curves for speech recognition WAC vs. real–time factor obtained by adjusting pruning thresholds from responses to *a)* **"greeting" and** *b)* **"card number" queries.**

4.3 Language Modeling with LBEs

The phrase–based bigram described in Section 4.1 was enhanced by including LBE symbols in the training transcriptions so that the language model was defined over the augmented dictionary $V \cup \Delta$. Overall, the number of acquired phrases for this model was 469 with phrase length varying in the range from one to twelve symbols. Examples of automatically acquired phrases include "`I would like to make a collect call`", "`a [wfrag]`", "`<dig3> [brth] <dig3>`", and "`[brth] and I`" where `<dig3>` is a non–terminal symbol representing a sequence of three natural numbers. This new language model has a perplexity of 16.0, a reduction of thirteen percent relative to the baseline phrase–based bigram in Section 4.1.

The ASR word accuracy for the LBE–based language model on both the greeting and card number test sets is given in the third row of Table 3. There is a significant increase in WAC for the more constrained "card number" utterances relative to the case where LBE symbols are included in the acoustic model alone. However, the word accuracies are identical for the greeting utterances. One possible explanation for this is that there is a considerably more regular "structure" to the card number utterances. This is illustrated by the fact that only one percent of the card number utterances contain out–of–vocabulary words (OOV's), where thirty percent of the greeting utterances contain OOV's.

5 SUMMARY AND DISCUSSION

An experimental study evaluating the effects on asymptotic ASR performance of incorporating representations of non–speech events in acoustic and language models has been presented. However, the impact of modeling these events is even more apparent when analyzing the ASR performance over a range of operating points. Figure 2 displays the WAC versus recognition time for each of the three systems given in Table 3 on the greeting and card number data sets. The WAC obtained using LBE–based acoustic and language models at CPU seconds per audio second equal to one is over six absolute percentage points greater than the baseline system on both test sets. It is also clear from Figure 2 that the real–time performance for the LBE–based system has come very close to the asymptotic ASR performance given in Table 3. This has demonstrated the practical importance of characterizing non–speech events for this task and incorporating them into ASR system design.

6 ACKNOWLEDGEMENTS

The authors would like to express their appreciation to Sandy Gates and the speech labeling lab at ATT Labs–Research for providing the transcriptions used in this work.

REFERENCES

[1] A.L.Gorin, G.Riccardi, and J.H.Wright. How may I help you? *Speech Communication*, 23:113–127, 1997.

[2] D. Liu, L. Nguyen, S. Matsoukas, J. Davenport, F. Kubala, and R.Schwartz. Improvements in spontaneous speech recognition. *Proc. DARPA Speech Recognition Workshop*, February 1998.

[3] G. Riccardi, A. L. Gorin, A. Ljolje, and M. Riley. A spoken language system for automated call routing. *Proc. Int. Conf. on Acoust., Speech, and Sig. Processing*, pages 1143–1146, April 1997.

[4] G. Riccardi, R. Pieraccini, and E. Bocchieri. Stochastic automata for language modeling. *Computer Speech and Language*, 10:265–293, 1996.

[5] A. Stolcke and E. Shriberg. Statistical language modeling for speech dysfluencies. *Proc. Int. Conf. on Acoust., Speech, and Sig. Processing*, pages 405–408, May 1996.