MODELLING THE PREPAUSAL LENGTHENING EFFECT FOR SPEECH RECOGNITION: A DYNAMIC BAYESIAN NETWORK APPROACH

Ning Ma[1*], *Chris D. Bartels*[2], *Jeff A. Bilmes*[2] *and Phil D. Green*[1]

[1]Department of Computer Science, University of Sheffield, Sheffield S1 4DP, UK
[2]Department of Electrical Engineering, University of Washington, Seattle, WA 98195
{n.ma,p.green}@dcs.shef.ac.uk, {bartels,bilmes}@ee.washington.edu

ABSTRACT

Speech has a property that the speech unit preceding a speech pause tends to lengthen. This work presents the use of a dynamic Bayesian network to model the prepausal lengthening effect for robust speech recognition. Specifically, we introduce two distributions to model inter-state transitions in prepausal and non-prepausal words, respectively. The selection of the transition distributions depends on a random variable whose value is influenced by whether a pause will appear between the current and the following word. Two experiments are presented here. The first one considers pauses hypothesised during speech decoding. The second one employs an extra component for speech/non-speech determination. By modelling the prepausal lengthening effect we achieve a 5.5% relative reduction in word error rate on the 500-word task of the SVitchboard corpus.

Index Terms— Prepausal lengthening, duration, prosody, robust speech recognition, dynamic Bayesian networks

1. INTRODUCTION

Automatic speech recognition (ASR) employing segmental features (e.g. MFCC) has achieved great success, but performance often degrades dramatically in the presence of noise. One reason is that most ASR systems do not explicitly represent prosodic properties such as duration. Modelling their interaction with words is important as prosodic properties can be relatively insensitive to moderate noise and channel distortions [1]. Their resistance to noise conditions also allows prosody analysis on the training data to be valid for ASR in a condition that is unknown to match the training condition. In this study we propose to model one prosodic property – the prepausal lengthening effect on word durations.

The prepausal lengthening effect is the property that before a speech pause, the preceding speech unit tends to lengthen. The nature and effects of this property has been well studied in [2, 3, 4] through a series of experiments analysing segmental durations in continuous speech. These studies have given evidence that the syntactic pause is one of the primary factors that influence vowel durations for an individual speaker. The lengthening property is thought to be correlated with high-level linguistic structures such as sentence boundaries, syntax and semantics, but it can also be observed in connected digits where most linguistic cues are minimised [5].

Since this duration property occurs in speech units such as phones, syllables and words, most research has focused on the use of phone-/word-level models for ASR. In [6] ASR improvements were reported by penalising word hypotheses that are inconsistent with prosodic duration. This idea was extended in [7] and [1] where explicit word-duration models were estimated and employed to re-score word hypotheses in N-best lists. To model prepausal lengthening, separate duration models for words preceding a pause were employed, which significantly reduced word errors. [8, 5] also reported ASR improvements by employing separate duration models for sentence-final words. Prepausal lengthening was also investigated in [9] within a hierarchical duration model framework, although the property was not explicitly modelled.

This paper proposes the use of a dynamic Bayesian network (DBN) to model the prepausal lengthening effect for speech recognition. Specifically, we introduce two state transition matrices for prepausal and non-prepausal words, respectively. The selection of the transition matrix depends on a random variable whose value is influenced by whether a pause will appear between the current and the following word. In this study the 500-word task of the SVitchboard corpus [10] is used, which is a small subset of Switchboard I [11] with closed vocabulary. In Section 2 we will explore the prepausal lengthening effect further using this corpus. Section 3 presents techniques to incorporate this property into ASR. Experiments and results will be described in Section 4. Section 5 concludes and presents future directions.

2. PREPAUSAL LENGTHENING IN SVITCHBOARD

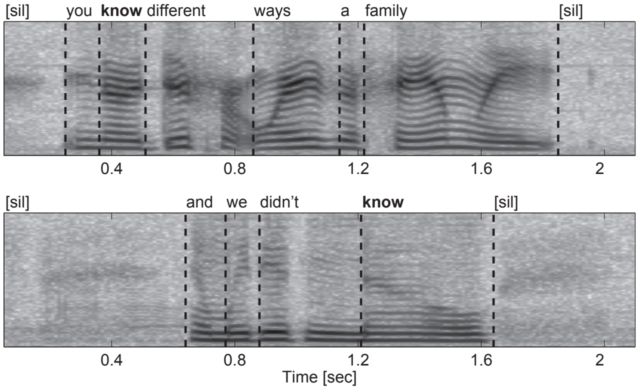

Fig. 1. An example from the SVitchboard corpus to illustrate the prepausal lengthening effect. The transcription is shown at the top of the spectrogram of each audio signal with segmentation indicated by dashed lines. The word $know$ lasts 141 ms in (a) and 436 ms in (b) where it precedes a speech pause ($[sil]$).

[*] The first author performed the work while visiting the University of Washington, Seattle.

The prepausal lengthening effect is very strong in the SVitchboard corpus. Although there is an intro/inter-speaker difference in the speaking rate, the duration of words (mainly vowels) is heavily influenced by the following pause. Fig. 1 illustrates this effect. Two sentences both containing the word $know$ are used here. In sentence (a) $know$ occurs before another word and its duration lasts 141 ms. In sentence (b) where $know$ precedes a speech pause, its duration is significantly longer (436 ms).

Table 1. Mean durations (Mn.) and standard deviations (s.d.), in ms, of the 10 most frequently occurring words in the 500-word task of SVitchboard. N = number of cases. Mn.Inc. = Increase in mean duration. Inc.% = Percent increase in mean duration.

	Non-prepausal			Prepausal				
word	N	Mn.	s.d.	N	Mn.	s.d.	Mn.Inc.	Inc.%
I	7633	129	74	748	260	127	131	101%
and	4348	272	157	1055	398	170	126	46%
you	4456	126	57	589	251	105	124	98%
oh	2915	249	146	1305	527	234	278	112%
that	2935	209	95	1153	287	108	78	37%
right	677	366	159	2987	407	126	41	11%
it	2428	137	69	907	186	81	49	36%
know	2065	172	86	1091	290	114	118	68%
to	2500	124	79	412	298	131	174	140%
that's	2488	258	76	198	354	124	96	37%
						Mn.	121	68%
						s.d.	68	41%

Table 2. Word duration statistics of the 10 words in SVitchboard which caused the baseline recogniser the most substitution errors.

	Non-prepausal			Prepausal				
word	N	Mn.	s.d.	N	Mn.	s.d.	Mn.Inc.	Inc.%
it	2428	137	69	907	186	81	49	36%
I	7633	129	74	748	260	127	131	101%
that	2935	209	95	1153	287	108	78	37%
to	2500	124	79	412	298	131	174	140%
you	4456	126	57	589	251	105	124	98%
is	1066	189	105	276	368	152	178	94%
a	2145	82	67	431	229	118	147	178%
oh	2915	249	146	1305	527	234	278	112%
know	2065	172	86	1091	290	114	118	68%
the	2042	119	81	546	277	135	158	133%
						Mn.	143	100%
						s.d.	62	45%

We analysed word duration using 55,504 sentences (about 33 hours long) from the SVitchboard corpus. The duration samples were obtained from Viterbi forced-alignments and were divided into two parts. Words followed by a pause longer than 200 ms are considered as prepausal, and the rest are considered as non-prepausal. The threshold was used to remove pauses required for articulation (i.e. the filled pause [12]). Since the typical duration of a syllable in English speech is around 200 ms, pauses are more easily perceived if their duration is longer than 200 ms [13]. Experiments reported in [9] also showed that pauses shorter than 200 ms do not significantly affect the duration of the preceding speech unit.

Table 1 presents word duration statistics of the 10 most frequently occurring words in the 500-word task of SVitchboard. The average increase in prepausal word duration is 121 ms, or 68% of the average non-prepausal duration. The lengthening property is strongly affected, however, by the presence/absence of a final consonant. Words ending with a consonant (e.g. $that$), which have an average duration increase of 78 ms, are much less lengthened by the following pause than words ending with a vowel (e.g. to), which have an average duration increase of 165 ms. This observation is consistent with studies reported in [3].

A strong lengthening effect is also observed with the words that caused the baseline recogniser (see Fig. 2) the most errors. Table 2 shows the duration statistics of the 10 words that caused the most substitutions. The average increase in prepausal word duration is 143 ms, or 100% of non-prepausal duration. A similar duration increase is found with the words that caused the most insertions and deletions. Identical duration analysis (not included here) that was run on the complete Switchboard I corpus also demonstrates a strong prepausal lengthening effect.

3. MODEL

The baseline system is a conventional hidden Markov model (HMM) implemented using the DBN shown in Fig. 2. This graph uses state-clustered within-word triphones and implements a three-state left-to-right topology. See [14] for more treatment on DBNs in speech recognition.

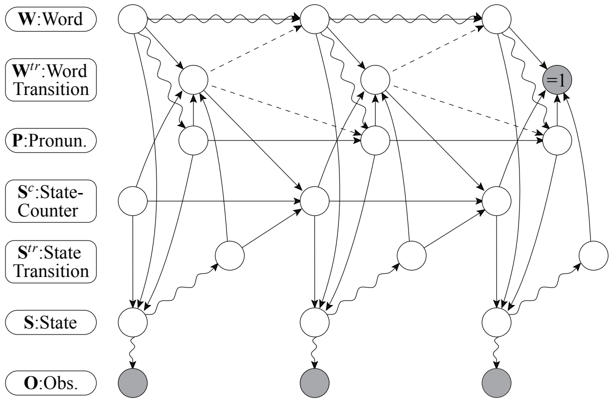
W:Word
$\mathbf{W}^{tr}$:Word Transition
P:Pronun.
$\mathbf{S}^{c}$:State-Counter
$\mathbf{S}^{tr}$:State Transition
S:State
O:Obs.
=1

Fig. 2. The baseline model [14] is a standard speech HMM represented as a DBN. Hidden variables are white while observed variables are shaded. Straight arrows represent deterministic relationships, curvy arrows represent probabilistic relationships, and dashed arrows are switching relationships.

The basic approach to modelling the prepausal lengthening effect is to lengthen a word by slowing down its inter-state transitions if a pause is considered to occur after. This model, as shown in Fig. 3, adds several additional components to the baseline model. The lower portion of the graph is mostly identical to the baseline model except that the variable *State Transition* can utilise different state transition matrices depending on context. The two transition matrices are used respectively for prepausal words and non-prepausal words, and they are learnt from the training data. Let $\mathcal{A}_{pp}$ represent the one for prepausal words and $\mathcal{A}_{np}$ represent the one for non-prepausal words. Since prepausal words are normally lengthened, self-transition probabilities in $\mathcal{A}_{pp}$ will be higher than in $\mathcal{A}_{np}$.

The selection of the state transition matrix depends on the switching parent of *State Transition*, a new variable *RelativeShort-*

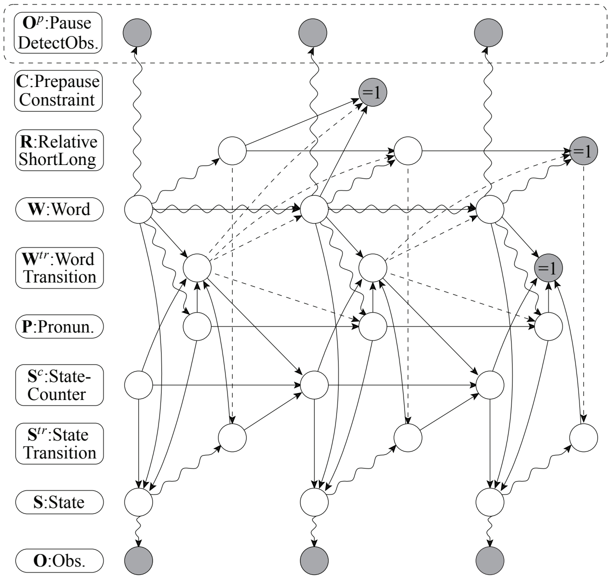

Fig. 3. A DBN graph modelling the prepausal lengthening effect (see Fig. 2 for key). The full model is called **Prepausal+PauseDetector**. The model without the variable *PauseDetectObs* at the top (dashed) is called **Prepausal**. See text for more details.

Long, notated $\mathbf{R}$. This random variable $\mathbf{R}$ can take a binary value: $short$ indicates that the current word is not being lengthened and transition matrix $\mathcal{A}_{np}$ should be used; $long$ indicates that the current word is being lengthened and $\mathcal{A}_{pp}$ should be used. Variable $\mathbf{R}$ itself has a switching parent, *Word Transition*. When there is no word transition, it simply copies its value in the last frame. If there is a word transition, $\mathbf{R}$ is governed by a distribution conditional on the variable *Word*, $\mathbf{W}$. The distribution $p(\mathbf{R}|\mathbf{W})$ is learnt from the training data.

The variable *Prepause Constraint* is a constraint that is active only when there is a word transition. It has conditional parents of *RelativeShortLong* of the last frame (notated $\mathbf{R}(-1)$) and *Word* of the current frame (notated $\mathbf{W}(0)$. When there is a word transition, this constraint enforces two rules:

1. $\mathbf{R(-1)} = short$ and $\mathbf{W(0)} \neq silence$, or
2. $\mathbf{R(-1)} = long$ and $\mathbf{W(0)} = silence$.

The constraint variable is always observed to be value 1. This will cause any decoding hypotheses that do not follow these rules to be eliminated (i.e. to ensure prepausal lengthening). When the constraint is inactive, it has no effect on the model.

The model with the components described so far is called **Prepausal** (i.e. the graph without the dashed variable at the top in Fig. 3). It considers word hypotheses formed during decoding for pause/non-pause determination. We also test a model that employs pause detecting features. These features are represented by an observed variable, *PauseDetectObs*, using two mixtures – one for all speech words and the other for the silence word. This model (i.e. the full graph in Fig. 3) is called **Prepausal+PauseDetector**. In this model, the variable *PauseDetectObs* directly affects the variable *Word*, and its weight influences the final results. The pause detecting features are modelled by using Gaussian mixtures with diagonal-covariance and details will be given in Section 4.

The Gaussian parameters trained for the baseline model were imported directly into the two new models. While training parameters of the new models, these baseline-model Gaussian parameters were held fixed. Three distributions that need to be trained for the 'Prepausal' model are: $\mathcal{A}_{np}$, $\mathcal{A}_{pp}$, and $p(\mathbf{R}|\mathbf{W})$. In the 'Prepausal+PauseDetector' model the extra Gaussian mixtures needed to model pause detecting features were also trained. The language model scale and word insertion penalty was determined by evaluating the recognition performance over a range of settings on the development set. The new models have an additional scaling factor on the transition distributions $\mathcal{A}_{np}$ and $\mathcal{A}_{pp}$. The 'Prepausal+PauseDetector' model also has a scaling factor on the pause detecting features. These scales along with the language model scale and word insertion penalty were optimised on the development set separately from the baseline.

4. EXPERIMENTS AND RESULTS

All experiments were performed on the 500-word task of the SVitchboard corpus [10]. The A, B, and C folds were used for training, the D_short fold was used as the development set, and the E fold was used as the evaluation set. The acoustic observation vectors consist of 13-dimensional perceptual linear prediction (PLP) features normalised on a per-conversation-side basis along with their deltas (D) and accelerations (A). The features are modelled by using 32-component Gaussian mixtures with diagonal-covariance. All models were trained and decoded using the Graphical Models Toolkit (GMTK) [15].

We tested various voice activity detection (VAD) features for the 'Prepausal+PauseDetector' model. This was done using a separate HMM-based pause detector similar to the one used in [16], which consists of an ergodic HMM with two states – $speech$ and $pause$. Table 3 lists the detection accuracy rates. The 'Energy' feature consists of energy and delta energy smoothed over a 9-frame Hamming window. 'VAD5' consists of 5 features commonly used for VAD: Energy, Energy Entropy, Zero-Crossing Rate, Spectral Roll-off, and Spectral Centroid. PLP features were also tested.

Table 3 shows that energy-based features perform better than those characterising speech (e.g. spectral centroid). This is because in SVitchboard the leakage of speech from the other channel may appear during silence periods. Many speech frames are falsely detected as pause. The energy-based features, however, are less affected by this problem. The lowest frame error rate is achieved using 'PLP_D' (10.3 for the development set). We employ this feature with the 'Prepausal+PauseDetector' model in ASR experiments.

Table 4 lists results of ASR experiments. Using the 'Prepausal' model we achieve a 5.5% relative reduction in word error rate (WER), which is significant at the 0.001 level (matched-pairs test).

Table 3. Frame error rates of speech/pause detection. *Dim* is the feature dimensionality and *Comp* is the number of mixture components used. All mixtures have diagonal covariance. The numbers in () indicate percent error rates of respective speech/pause frames.

			Frame Error Rate (%)	
Feature	Dim	Comp	Development	Evaluation
Energy	2	2	14.3 (14.5/14.1)	13.7 (13.4/14.0)
VAD5	5	8	15.7 (19.8/12.4)	15.4 (20.1/11.2)
PLP	13	32	13.3 (17.3/9.9)	13.7 (17.6/10.3)
PLP_D	26	32	**10.3** (9.3/11.1)	**10.6** (9.6/11.5)
PLP_D_A	39	32	12.0 (12.8/11.0)	12.5 (13.0/12.0)

Table 4. Speech recognition results on the 500-word task of SVitchboard. *S*, *D*, and *I* are counts of substitutions, deletions, and insertions.

	Development				Evaluation			
Model	S	D	I	WER	S	D	I	WER
Baseline	602	190	197	53.9%	7069	2634	2336	60.1%
Prepausal	584	223	129	51.0%	6937	3035	1408	**56.8%**
Prepausal+PauseDetector	583	230	139	51.9%	6969	2983	1506	57.2%

The improvement is mainly from reducing insertions (by 830), due to the fact that words preceding a pause are hypothesised as lengthened in the model. It is also likely that lengthened prepausal words match the acoustics better as substitutions are also reduced (by 100). The model produces more deletions (by 349). Using the 'Prepausal+PauseDetector' model we achieve a 4.8% relative reduction, and the improvement is significant at the 0.001 level.

The difference between results of the two prepausal models is not significant. We believe this is because there already is a state for pause in the base model, and that information about this state is effectively being communicated from the existing speech features (PLPs) via the phone variable (which is set to the pause state) to the silence word (the silence word should be the one that best explains a set of phones being in the pause state for a duration $>$ 200 ms). The PLPs used for speech/non-speech detection might thus be redundant with the normal speech features. It may be that the initial speech/pause detection analysis selected the best features only for this subtask rather than the ones that would work best in the final combined model. Future work will investigate this hypothesis and will employ secondary features in novel ways.

Table 5. Frame error rate (%) of speech/pause segmentation produced by various models on the development set.

Baseline	Prepausal	Prepausal+PauseDetector
12.7 (4.5/19.4)	8.5 (11.3/6.2)	8.5 (11.0/6.5)

Table 5 gives speech/pause segmentation error rates produced by various decoders. This is done by comparing the decoder output with reference pause segmentation (from forced-alignments) on a frame-by-frame basis. It is clear that with the baseline model many more pause frames are falsely recognised as speech, causing many insertion errors. With the two proposed models the pause segmentation error rate is greatly reduced. In fact, the results are better than any pause detectors reported in Table 3. In the 'Prepausal+PauseDetector' model the extra pause detection component does not bring much gain since the error rates of pause segmentation output by both prepausal models are almost identical.

5. CONCLUSIONS

In this study we investigate the prepausal lengthening effect and incorporate this property into speech recognition using a dynamic Bayesian network. The lengthening effect is very strong in conversational speech and by modelling the property we achieve a significant reduction in WER. It has been shown [3] that speech pauses affect the preceding vowels more than consonants. Currently phones are lengthened regardless of their categories. This could potentially be improved upon by sharing transition probabilities for phones with less lengthening effect. Another property that is not modelled here is that speech acoustics also change due to the lengthening effect. Future work will investigate if this has a significant impact on ASR.

Acknowledgements: This work was partly supported by the Worldwide Universities Network – 'Research Mobility Programme' and ONR MURI grant N000140510388.

6. REFERENCES

[1] D. Vergyri, A. Stolcke, V. Gadde, L. Ferrer, and E. Shriberg, "Prosodic knowledge sources for automatic speech recognition," in *Proc. IEEE ICASSP*, 2003, pp. 208–211.

[2] D. Klatt, "Linguistic uses of segmental duration in English: Acoustic and perceptual evidence," *J. Acoust. Soc. Am.*, vol. 59, no. 5, pp. 1208–1221, 1976.

[3] T. Crystal and A. House, "Segmental durations in connected-speech signals: Syllabic stress," *J. Acoust. Soc. Am.*, vol. 83, no. 4, pp. 1574–1585, 1988.

[4] W. Campbell, "Segment durations in a syllable frame," *Journal of Phonetics*, vol. 19, pp. 37–47, 1991.

[5] N. Ma, J. Barker, and P. Green, "Applying duration constraints by using unrolled HMMs," in *Proc. Interspeech*, Antwerp, 2007, pp. 1066–1069.

[6] A. Stolcke, E. Shriberg, D. Hakkani-Tür, and G. Tür, "Modeling the prosody of hidden events for improved word recognition," in *Proc. Eurospeech*, Budapest, 1999, pp. 307–310.

[7] V. Gadde, "Modeling word durations," in *Proc. ICSLP*, Beijing, 2000, pp. 601–604.

[8] N. Ma and P. Green, "Context-dependent word duration modelling for robust speech recognition," in *Proc. Interspeech*, Lisbon, 2005, pp. 2609–2612.

[9] G. Chung and S. Seneff, "A hierarchical duration model for speech recognition based on the ANGIE framework," *Speech Commun.*, vol. 27, pp. 113–134, 1999.

[10] S. King, C. Bartels, and J. Bilmes, "SVitchboard 1: Small vocabulary tasks from Switchboard 1," in *Proc. Interspeech*, Lisbon, 2005, pp. 3385–3388.

[11] J. Godfrey, E. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *Proc. IEEE ICASSP*, San Francisco, CA, 1992, vol. 1, pp. 517–520.

[12] B. Zellner, "Pauses and the temporal structure of speech," in *Fundamentals of speech synthesis and speech recognition*, E. Keller, Ed., pp. 41–62. Chichester: John Wiley, 1994.

[13] F. Goldman-Eisler, *Psycholinguistics: Experiments in spontaneous speech*, New York: Academic Press, 1968.

[14] J. Bilmes and C. Bartels, "A review of graphical model architectures for speech recognition," *IEEE Signal Processing Magazine*, vol. 22, no. 5, pp. 89–100, 2005.

[15] J. Bilmes, *GMTK: The Graphical Models Toolkit*, 2002.

[16] A. Acero, C. Crespo, and J. Torrecilla, "Robust HMM-based endpoint detector," in *Proc. Eurospeech*, Berlin, 1993, pp. 1551–1554.