Applying word duration constraints by using unrolled HMMs

Ning Ma, Jon Barker, Phil Green

Speech and Hearing Research Group, Department of Computer Science,
University of Sheffield, UK
{n.ma, j.barker, p.green}@dcs.shef.ac.uk

10.21437/Interspeech.2007-105

Abstract

Conventional HMMs have weak duration constraints. In noisy conditions, the mismatch between corrupted speech signals and models trained on clean speech may cause the decoder to produce word matches with unrealistic durations. This paper presents a simple way to incorporate word duration constraints by unrolling HMMs to form a lattice where word duration probabilities can be applied directly to state transitions. The expanded HMMs are compatible with conventional Viterbi decoding. Experiments on connected-digit recognition show that when using explicit duration constraints the decoder generates word matches with more reasonable durations, and word error rates are significantly reduced across a broad range of noise conditions.

1. Introduction

Automatic speech recognition (ASR) based on hidden Markov models (HMMs) has achieved great success, but performance often degrades significantly in the presence of noise. One reason is that conventional HMMs have weak duration models. The state duration distribution is implicitly modelled by an inappropriate Geometric distribution and there is no modelling of word durations [1]. Therefore the process of decoding noise-corrupted speech may produce word matches with unrealistic durations, given models trained on clean speech. This sometimes has disastrous consequences during the matching process: word strings where the associated word models have short durations tend to be favoured over competing strings with fewer words but longer durations. This effect can be observed in a connected-digit recognition task with no grammar constraints, where the number of insertion errors greatly exceeds that of deletions and substitutions in noisy conditions [2].

There have been several attempts to model explicitly state durations by adapting HMM-based systems (e.g., [3] and [4]). However, the minor improvements produced often do not justify the extra complexity introduced [5]. While the meaning of modelling state-level durations is obscure, modelling word-level duration constraints is potentially more effective for improving ASR in noise [6]. Word durations are, despite of the influence of the Lombard effect on speech [7], relatively insensitive to moderate noise levels. However, with the Markov state independence assumption, modelling state duration does not necessarily produce a good model of word durations. Ma and Green [6] have shown that an explicit word duration model can help the decoder to combat the corruption of acoustic features in noisy conditions.

This paper reports a generic way to employ word-level durational knowledge for robust speech recognition by unrolling HMMs to form a lattice where word duration probabilities can be applied directly to state transitions. The expanded HMMs are fully compatible with conventional Viterbi decoding. In the next section we investigate some characteristics of word durations. Section 3 presents techniques for expanding HMMs to implement the duration model. Recognition experiments using the Aurora 2 corpus are described and discussed in Section 4. Section 5 concludes and presents future directions.

This work was funded by UK EPSRC grant GR/R47400/01.

2. Word duration analysing and modelling

Crystal and House [8] performed a series of experiments analysing segmental durations in connected-speech signals, in an effort to apply durational information to the automatic analysis of speech. Among many factors that may influence segmental durations for an individual speaker, the stress patterns of a language are a primary factor. Speakers tend to lengthen syllables (or words) when stressing them. For example, in the Crystal and House experiments the mean duration of stressed vowels is found to be 70 ms greater than the average for unstressed vowels. Crystal and House also discussed a strong prepausal lengthening effect [9] on vowel durations, an effect in which vowels followed by syntactic pauses (e.g., sentence markers) are longer than the others. In a connected-digit domain where high-level linguistic cues are minimised, this effect can also be observed [6].

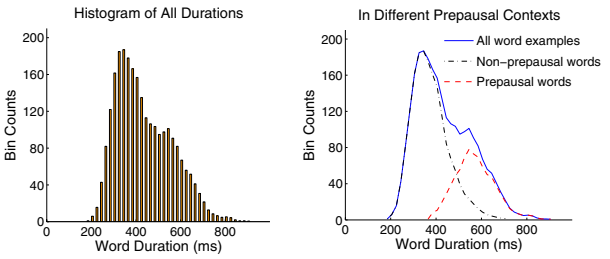

Figure 1: Word duration histograms for digit 'six'. Left: raw histogram computed from all duration examples. Right: a comparison of duration histograms in different contexts.

Column 'All' in Table 1 shows durations of various digits in the Aurora 2 corpus in milliseconds. These duration examples were obtained from the Aurora training data using an automatic Viterbi alignment for each word in the vocabulary based on a set of well-trained word-level HMMs. Different words have different duration statistics. They are difficult to model accurately with a single Gaussian because their distribution has a skewed shape, as shown in the left panel of Fig. 1. As word durations are themselves discrete, a discrete distribution is attractive for a small vocabulary task where sufficient training data are avail-

Table 1: *Mean durations (Mn.) and standard deviations (s.d.), in milliseconds, of various digits in Aurora 2 corpus. Prepausal context as indicated. N = number of cases; Mn. Inc. = relative mean duration increase in the prepausal context.*

	All examples			Non-prepausal			Prepausal			
word	N	Mn.	s.d.	N	**Mn.**	s.d.	N	**Mn.**	s.d.	**Mn. Inc.**
one	2545	357	94	1784	325	80	761	432	80	33%
two	2531	338	99	1757	302	85	774	421	75	40%
three	2521	349	92	1769	314	75	752	433	72	38%
four	2539	373	97	1748	338	83	791	450	79	33%
five	2491	407	111	1698	360	83	793	509	96	41%
six	2545	436	123	1756	374	79	789	572	87	53%
seven	2525	426	88	1778	397	77	747	493	74	24%
eight	2515	323	95	1752	280	64	763	420	83	50%
nine	2492	406	99	1715	369	80	777	488	87	32%
oh	2500	324	94	1769	291	79	731	402	80	38%
zero	2523	448	100	1761	419	91	762	515	88	23%

able. We therefore employ a histogram-based word duration model proposed in [6]. This word duration model will not scale to medium or large vocabulary tasks because it may become intractable to collect robust word duration statistics. Therefore a parametric model (e.g., a Gaussian mixture model) may be employed[1]. The duration examples are used to compute histograms with a bin width of 10 ms. Let $P(d|w)$ denote the probability of word w having a duration d. To evaluate $P(d|w)$, the histograms are smoothed using a 5-point median filter and then normalised to have area 1. Because of the high dimensionality of the feature vectors typically used, there is a scaling factor to control the impact on the decoding procedure, forming the word duration penalty $D(d|w)$:

$$D(d|w) = P(d|w)^{\gamma} \tag{1}$$

where γ is the empirical scaling factor on word durations.

The left panel in Fig. 1 shows that the duration histogram for digit 'six' has a bimodal distribution. This observation is due to the prepausal lengthening effect discussed previously. To further examine this effect we divide the duration examples of each digit for two conditions: examples followed by a digit and examples preceding a long pause. In Aurora 2 corpus there is a long pause at the end of each utterance and our experiments show that the brief inter-digit pauses in some long digit strings do not give a strong prepausal lengthening effect. Therefore only the sentence-final words are considered as prepausal words. Two duration histograms are computed for the two conditions and their distributions are shown in the right panel of Fig. 1, along with the duration distribution of all examples. It is clear the bimodal distribution becomes two unimodal ones. Table 1 lists the mean and standard deviations of the duration of each digit in both the prepausal and non-prepausal contexts. Prepausal words demonstrate longer durations and the relative mean duration increase is up to 53% (digit 'six'). The duration standard deviations in each context are also narrower than those of all examples. The prepausal lengthening effect is observed for all the digits but is least strong for two-syllable digits ('seven' and 'zero'). This suggests that only the emphasised syllable may be lengthened but no further investigation has been done in this study. To model the prepausal lengthening effect, we estimate $P(d|w, c)$ – the probability of word w having a duration d in context c. In our case $c = (\text{prepausal} \mid \text{non-prepausal})$. By applying a scaling factor, we can compute the context-dependent word duration penalty:

$$D(d|w, c) = P(d|w, c)^{\gamma} \tag{2}$$

[1] A parametric model can be employed in the same manner.

3. Incorporating duration constraints

With standard HMMs we wish to apply word duration constraints to word sequence hypotheses as they leave word-final states. This cannot be done directly in the Viterbi algorithm as it does not keep a record of durations of different paths. Therefore it is not possible to correctly apply duration penalties. Ma and Green [6] proposed a multi-stack decoding algorithm to incorporate a word duration model based on the NOWAY decoder [10]. It extracts the most likely hypothesis from every stack, computes one-word extensions, applies word duration penalties for the word, and places all the new hypotheses into corresponding stacks. However, most HMM-based ASR systems employ a Viterbi decoder and it may not always be feasible to incorporate a stack decoder. In this section we propose a more generic technique which amounts to little more than expanding the HMM topologies so that word duration penalties can be incorporated. This technique is fully compatible with existing ASR systems. It is theoretically equivalent to the multi-stack decoding technique and therefore produces the same recognition results as reported in Section 4.2.

Assuming a no-skip, left-to-right HMM with N states $q_1, q_2, \ldots, q_N$ for word w is being expanded. We can compute corresponding duration penalties using Eq. (1) with an expected duration range d^w_{min} to d^w_{max}. Each emitting state q_i is then duplicated d^w_{max} times, which share the same Gaussian parameters of q_i. The duplicated states q_i form a sequence and the self-transition of each state is replaced by a one-way transition between two adjacent states. For N states in the old HMM we get N state sequences. Except for the last state sequence q_N, the j^{th} state in the sequence q_i is connected to the $(j+1)^{th}$ state in the next state sequence q_{i+1}, with the transition probability the same as that from q_i to q_{i+1} in the old HMM. The beginning non-emitting state is connected to the first state in the first state sequence q_1 with a transition probability of 1.0 as an entry point. Each state in the last sequence is connected to the non-emitting state at the end as one of the terminating points in the expanded model. When word sequence hypotheses leave a model from any of the final states q_N, different paths within the model to the leaving state are guaranteed to have the same duration. Therefore the duration penalty can be safely applied in the expanded HMM. We use the corresponding duration penalty

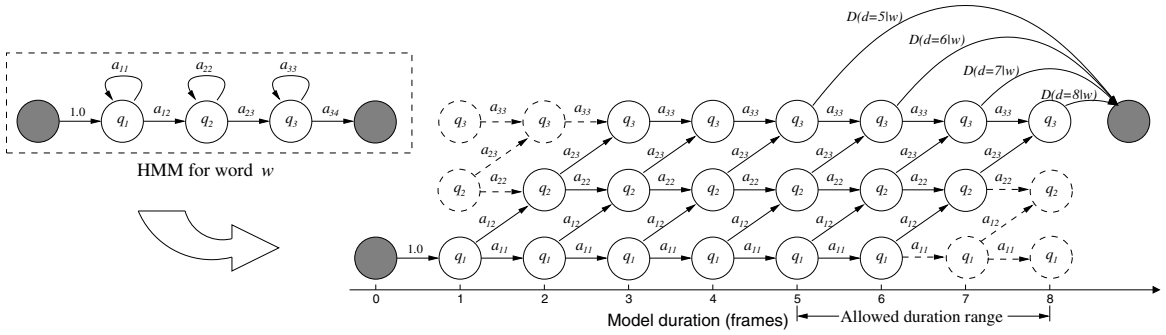

Figure 2: Illustration of unrolling a standard no-skip, left-to-right HMM with word duration penalties.

to replace the transition probability from each state in the last sequence q_N to the terminating non-emitting state.

Fig. 2 illustrates this procedure using an example of a 3-state no-skip, left-right HMM with 2 non-emitting states (dark circles) at the two ends. The states that will not be visited are marked with dashed circles. To make the expanded HMMs more efficient, we do not supply the terminating transition to the states before the $(d^w_{min})^{th}$ state in the last state sequence. In this example the allowed duration range for word w is 5 to 8 frames, therefore there are no such transition for the first four q_3 states. The allowed duration ranges are determined by examining the duration statistics obtained from the training data. A typical duration range for a non-prepausal digit in the Aurora corpus is 200 – 700 ms and for a prepausal digit a typical duration range is 300 – 900 ms. With a 10 ms frame shift although the state space is expanded roughly by a factor of 90, the computational load increases only by a much smaller factor in a small vocabulary task. Most the computation is for the observation probabilities, which remains constant because the Gaussian mixtures are tied up.

For word duration modelling in the prepausal context, we expand two sets of HMMs using Eq. (2): *NPPdigit* – HMMs for non-prepausal digits; and *PPdigit* – HMMs for prepausal digits. The decoder simply employs the two model sets with an EBNF grammar as follows: (sil {$NPPdigit} ($PPdigit) sil).

4. Experiments and results

4.1. Recognition systems

The experiments reported here employ the Aurora 2 speaker independent connected-digit recognition task. Spectral features were used so that missing data techniques can be applied [11]. Feature vectors were obtained via a 32-channel Gammatone filterbank distributed in frequency between 50 Hz and 3850 Hz on the equivalent rectangular bandwidth (ERB) scale. The features were supplemented with their temporal derivatives to form a 64-dimensional feature vector. Gender-dependent word-level HMMs were trained on the Aurora clean speech training set. Digit models ('1'–'9', 'oh' and 'zero') consist of 16 no-skip, left-right states with observations modelled by 7-component diagonal GMMs. A 3-state silence model was used to model the long pauses before and after an utterance and an additional 1-state silence model was used to model the brief inter-digit pauses that may occur during long digit strings.

The recognition system is a 'missing data' recogniser. The 'missing data' approach [11] assumes that when the speech is one of several sound sources, some spectro-temporal regions will remain uncorrupted and can be used as reliable evidence for recognition. The uncorrupted regions can be labelled using a spectrographic mask. The baseline system the best performing system described in [12], which takes combined harmonicity and SNR-based masks and uses gender dependent modelling. The second recognition system employs HMMs expanded with the pause-context-free word duration penalties calculated using Eq. (1) (NPP-WD system). The last recognition system uses the two sets of HMMs unrolled with Eq. (2) (PP-WD system). The scaling factor γ was set to 10 for all noise conditions in this study, which was tuned based on a small set of developing data.

4.2. Results and discussion

Fig. 3 shows the average word error rates (WER) over the four different noise types in Aurora 2 test set A at various SNR levels. The systems with both duration models clearly outperform the baseline system at low SNR levels. No significantly different results are achieved at SNRs above 15 dB as the baseline system is capable of achieving high performance in quiet conditions in this task. In noisy conditions the estimation of missing data masks becomes more difficult. When more data is missing duration constraints become more important. This is analogous to increasing the contribution of the language model when the acoustic model is poor.

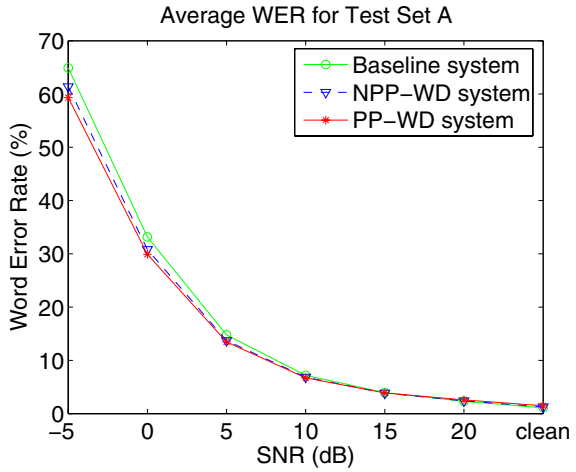

Figure 3: Average word error rates for test set A in Aurora 2 corpus at various SNR levels.

The PP-WD system achieves the lowest WERs, but the re-

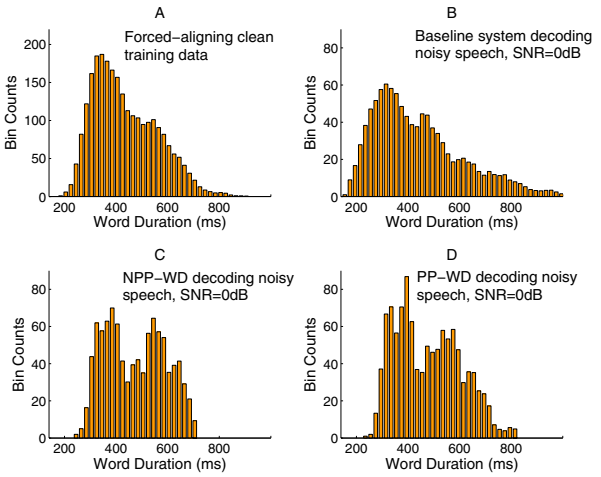

Figure 4: A comparison of word duration histograms of digit 'six' obtained from recognition results of various systems.

sults of the NPP-WD system are close. This is because without considering the pause context the 'prepausal' portion of the duration distribution is still well modelled by the histogram-based model. The performance gain using the prepausal context is mainly due to the emphasis of the prepausal duration distribution and a better estimate of the allowed duration ranges.

To examine the impact of the word duration constraints on the recogniser, we also compared the duration statistics produced during the decoding process. Word duration examples were collected in the back-tracing stage of various recognition systems. Histograms of these duration examples are then computed and compared to those obtained by forced-aligning the training data. Fig. 4 shows these duration histograms for digit 'six' at the SNR level of 0 dB. Panel (A) is the duration histogram obtained from forced-aligning the clean Aurora training data (same as in Fig. 1). Panels (B–D) show the histograms produced by the baseline recogniser, the NPP-WD system and the PP-WD system, respectively. When decoding noisy speech, the baseline recogniser generates many word matches with too short or too long durations and fails to demonstrate the second peak in the distribution around 572 ms. The proposed word duration model forces the recogniser to focus on word matches with more realistic durations and with the prepausal context it produces a duration distribution more similar to that from training data.

The examples used to estimate the word duration penalties in this study are only approximate digit durations as they are obtained by forced-aligning the training data with trained conventional HMMs, which do not encode correct duration models in the first place. This is generally acceptable in this digit recognition task as with clean training data the HMMs give reasonable duration statistics. Note that the 16-state no-skip, left-right topology forces the minimum duration of any word matches produced by the decoder to be 160 ms. Our preliminary experiments show that in noisy conditions using 16-state HMMs the decoder gave significantly lower WERs than using HMMs with fewer states, even the total numbers of parameters are same (i.e., the HMMs with fewer states are given more Gaussian mixtures). One possible reason is that with fewer states the decoder may produce more word matches with durations shorter than 160 ms, resulting in more recognition errors. Experiments [6] have also shown that the allowed duration ranges themselves can help achieve lower WERs as hypothesis paths with unrealistic durations will be pruned out of the search.

5. Conclusions

Word durations are relatively insensitive to moderate noise levels. In this paper we present a generic method to explicitly employ word duration constraints in different pause contexts by using unrolled HMMs. The technique enables an existing ASR system to employ a word duration model when available. Experiments show that the technique is able to offer significantly lower WERs over a strong missing data baseline system in noisy situations. The system presented assumes that word durations remain constant in various noise conditions. Future work will include adapting the duration model based on speaking rates.

6. References

[1] L.R. Rabiner, J.G Wilpon, and F.K. Soong, "High performance connected digit recognition using hidden Markov models," *IEEE Trans. Acoust. Speech Signal Processing*, vol. ASSP-37, no. 8, pp. 1214–1225, Aug. 1989.

[2] K. Power, "Durational modelling for improved connected digit recognition," in *Proc. ICSLP*, Philadelphia, USA, 1996, pp. 885–888.

[3] M.J. Russell and R.K. Moore, "Explicit modelling of state occupancy in hidden markov models for automatic speech recognition," in *Proc. ICASSP*, 1985, pp. 5–8.

[4] D. Burshtein, "Robust parametric modeling of durations in hidden Markov models," in *Proc. ICASSP*, 1995, pp. 548–551.

[5] C. Mitchell, M. Harper, and L. Jamieson, "On the complexity of explicit duration HMMs," *IEEE Trans. Speech and Audio Processing*, vol. 3, no. 3, pp. 213–217, 1995.

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

[7] J. Junqua, "The Lombard reflex and its role on human listeners and automatic speech recognizers," *J. Acoust. Soc. Amer.*, vol. 93, no. 1, pp. 510–524, 1993.

[8] T.H. Crystal and A.S. House, "Segmental durations in connected-speech signals: Current results," *J. Acoust. Soc. Amer.*, vol. 83, no. 4, pp. 1553–1573, 1988.

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

[10] S. Renals and M. Hochberg, "Efficient evaluation of the LVCSR search space using the NOWAY decoder," in *Proc. ICASSP*, Atlanta, 1996, pp. 149–152.

[11] M.P. Cooke, P.D. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and uncertain acoustic data," *Speech Comm.*, vol. 34, pp. 267–285, 2001.

[12] J.P. Barker, M.P. Cooke, and P.D. Green, "Robust ASR based on clean speech models: an evaluation of missing data techniques for connected digit recognition in noise," in *Proc. Eurospeech*, Aalborg, Denmark, 2001, pp. 213–216.