MODELING LINGUISTIC SEGMENT AND TURN BOUNDARIES FOR N-BEST RESCORING OF SPONTANEOUS SPEECH

Andreas Stolcke

Speech Technology and Research Laboratory
SRI International, Menlo Park, CA, U.S.A.
http://www.speech.sri.com/
stolcke@speech.sri.com

ABSTRACT

Language modeling, especially for spontaneous speech, often suffers from a mismatch of utterance segmentations between training and test conditions. In particular, training often uses linguistically-based segments, whereas testing occurs on acoustically determined segments, resulting in degraded performance. We present an N-best rescoring algorithm that removes the effect of segmentation mismatch. Furthermore, we show that explicit language modeling of hidden linguistic segment boundaries is improved by including turn-boundary events in the model.

1. THE SEGMENTATION PROBLEM IN LANGUAGE MODELING

One of the problems encountered in speech recognition on continuous, spontaneous speech is the segmentation of long waveforms. Because current recognizers prefer short waveform segments for best performance and to limit computational resources, conversation-length waveforms are typically pre-segmented using simple acoustic criteria, such as locations of long pauses and turn switches. This creates several problems for language modeling:

- The segmentation algorithm used (including its parameters) influences the statistics embodied in the language model (LM), creating a potential mismatch between training and test set. Strictly speaking, one would have to resegment the training data, recreate the word-level transcriptions, and retrain the language model every time the segmentation process is modified.
- The acoustic segmentation typically yields units that are not linguistically coherent, and hence sub-optimal for language modeling. Language modeling research on spontaneous speech [10] shows that N-gram LMs based on complete utterance units give lower perplexity than those based only on acoustic segmentations. Furthermore, work reported in [12] showed that the word error rate on spontaneous speech can be reduced simply by resegmenting the speech at linguistic boundaries and using a language model based on the same segmentation.
- Explicit modeling of spontaneous speech phenomena such as disfluencies also requires modeling of linguistic (as opposed to acoustic) segment boundaries [15]. Similarly, sophisticated LMs modeling syntactic structure typically assume complete sentences as their input [12].

The following excerpt from the Switchboard corpus [2] illustrates the discrepancies between acoustic and linguistic segmentations. Linguistic segment boundaries are marked by <s>, whereas acoustic boundaries are indicated by //. A subset of acoustic boundaries corresponds to turn boundaries, indicated by <t>.

```
B: <t> Worried that they're not
going to get enough attention?
<s> //
A: <t> Yeah <s> and, uh, you
know, colds and things like that
get -- //
B: <t> Yeah.  <s> //
A: <t> -- spread real easy and
things, <s> but, and they're
expensive <s>
and, // course, // there's a lot
of different types of day care
available, too, // you know,
where they teach them academic
things.  <s> //
B: <t> Yes.  <s>//
```

As can be seen, linguistic and acoustic boundaries differ widely. Notice in particular how linguistic segments can run across turn boundaries.

2. HIDDEN SEGMENTATION MODELING

To overcome a segmentation mismatch between training and test conditions we can model segment boundaries as *hidden events* in the language model. We ignore the overt segment boundaries in the test material, and compute the probability of a word sequence assuming that non-overt segment boundaries (e.g., sentence boundaries <s>) can occur between any two words. Computationally, this is achieved by associating a hidden state (S or NO-S) with each word, corresponding respectively to the presence or non-presence of a hidden segment-boundary immediately preceding the word. If the language model is Markovian, as in the case of an N-gram model, we obtain a hidden Markov model, and the total sentence probability can be computed by a forward algorithm [9]. A corresponding

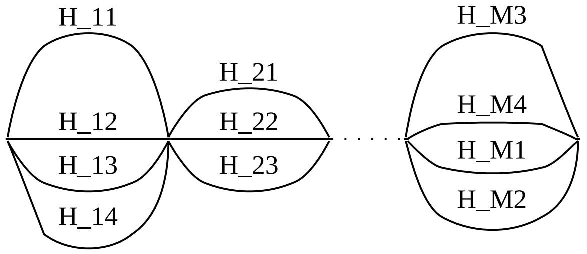

Figure 1. Lattice incorporating N-best hypotheses for all segments.

Viterbi algorithm can be used to find the most likely sequence of hidden S and NO-S states, corresponding to the most likely segmentation of a word sequence according to the language model. This is the basis of a simple automatic segmentation algorithm, and can been used to segment spontaneous speech transcripts into linguistic utterance units where hand-segmented transcripts are not available [14].

An approximate version of the hidden segmentation language model that does not require the forward algorithm has been used previously to study the effect of segmentation on language model perplexity [5].

3. N-BEST LIST RESCORING

To apply the hidden segmentation model to the output of a speech recognizer, we first generate N-best lists for each of the acoustic segments $X_1, X_2, \ldots, X_M$. These N-best lists are generated using a standard language model operating on one acoustic segment at a time. We based our implementation of the algorithm on N-best lists [8], but with minor changes the same methods could be applied to the rescoring of word lattices.

Let H_{ij}, $j = 1, \ldots, N$, be the N best hypotheses for the ith acoustic segment. The standard N-best scoring algorithm considers each acoustic segment in isolation:

- **Standard rescoring:** For each segment $i = 1, \ldots, M$, find the hypothesis that has highest posterior probability based on X_i alone:

$$\begin{aligned}&\operatorname*{argmax}_{H_{ij}} P(H_{ij}|X_i)\\ &= \operatorname*{argmax}_{H_{ij}} \frac{P_{\text{LM}}(H_{ij})P_{\text{AC}}(X_i|H_{ij})}{P(X_i)}\\ &= \operatorname*{argmax}_{H_{ij}} P_{\text{LM}}(H_{ij})P_{\text{AC}}(X_i|H_{ij})\end{aligned}$$

  P_{LM} is a standard language model operating on individual segments; P_{AC} is the acoustic model.

The goal of the new rescoring algorithms is to take the combined N-best hypotheses into account when choosing the best hypothesis for each segment, by applying the language model across acoustic segment boundaries. To this end, we combine all H_{ij} into a lattice (Figure 1) representing all possible combined hypotheses for $X_1, X_2, \ldots, X_M$. On this lattice, rescoring is performed using one of two dynamic programming methods:

- **Viterbi rescoring:** Find the sequence of hypotheses $H_{ij_i^*}$ and segmentations $s_{ij_i^*}$ that gives the highest overall probability under the combined acoustic model and the hidden segmentation language model. Let $H_s = \{H_{ij_i^*}, s_{ij_i^*}, i = 1, \ldots, M\}$ be the combined hypothesis including its segmentation, and let $X = (X_1, \ldots, X_M)$ be the combined acoustics. Viterbi rescoring finds

$$\begin{aligned}&\operatorname*{argmax}_{H_s} P(H_s|X)\\ &= \operatorname*{argmax}_{H_s} \frac{P_{\text{LM}}(H_s)P_{\text{AC}}(X|H_s)}{P(X)}\\ &= \operatorname*{argmax}_{H_s} P_{\text{LM}}(H_s)P_{\text{AC}}(X|H_s)\end{aligned}$$

  Here P_{LM} is the hidden segmentation language model; it operates on the combined hypothesis H_s irrespective of the acoustic segment boundaries between the H_{ij}.

- **Forward-backward rescoring:** For each acoustic segment X_i, find the hypothesis that has the highest posterior probability given the acoustics of the *entire* conversation. That is, for all $i = 1, \ldots, M$:

$$\begin{aligned}&\operatorname*{argmax}_{H_{ij}} P(H_{ij}|X)\\ &= \operatorname*{argmax}_{H_{ij}} \sum_{H_s : H_{ij} \in H_s} P(H_s|X)\\ &= \operatorname*{argmax}_{H_{ij}} \sum_{H_s : H_{ij} \in H_s} \frac{P_{\text{LM}}(H_s)P_{\text{AC}}(X|H_s)}{P(X)}\\ &= \operatorname*{argmax}_{H_{ij}} \sum_{H_s : H_{ij} \in H_s} P_{\text{LM}}(H_s)P_{\text{AC}}(X|H_s)\end{aligned}$$

  (The summation is over all combined hypotheses H_s that have H_{ij} as the part corresponding to the ith segment.)

Forward-backward rescoring is theoretically the better way of minimizing the per-segment error. This is because word error is additive over segments, and the error on each segment hypothesis is minimized by maximizing its posterior probability of correctness [13]. Viterbi rescoring is somewhat simpler and cheaper computationally; it also computes a best segmentation for the chosen hypotheses.

The acoustic models used in both algorithms are the same as in standard N-best rescoring. This is because current acoustic models exhibit no long-range dependencies, so that

$$P_{\text{AC}}(X|H_s) = \prod_{i=1}^{M} P_{\text{AC}}(X_i|H_{ij_i^*})$$

In the future, however, acoustic models could incorporate global characteristics of speech, such as speaking mode [7], in which case they, too, should operate across segment boundaries.

4. EXPERIMENTS

4.1. Data and Language Models

We tested the concepts and algorithms described here using the Switchboard corpus of spontaneous conversational speech [2]. We trained standard trigram language models with backoff smoothing [3] on 1.8 million words from that corpus. Two segmentations of the training data were used. In one case, the full training corpus was acoustically segmented, placing segment boundaries at turn boundaries and at pauses of at least 0.5 seconds. A second segmentation was available from the Linguistic Data Consortium, consisting of transcripts with hand-annotated linguistic segment boundaries [4].

However, only 1.4 million words were available in hand-segmented form. We therefore trained an automatic linguistic segmenter on this data, and used it to segment the remaining training data. This method had previously been shown to give good segment boundary detection accuracy on this corpus (85% recall, 3% false alarms) [14]. The hand-segmented and the automatically segmented training data were pooled, resulting in a linguistic segment language model based on the same amount of training data as the acoustic segment language model.

The test set consisted of 25 Switchboard conversations (24,000 words) and was acoustically segmented. For each segment, an N-best list of up to 2000 hypotheses was generated, using SRI's Decipher(TM) recognizer with continuous density genonic HMM acoustic models [1, 6]. For rescoring purposes, each side in a Switchboard conversation was treated as one stream of speech to be recognized, separate from the opposite side.

4.2. Complexity Issues

The time complexity of the dynamic programming algorithm scales with the square of the number of N-best hypotheses. To keep the computation in reasonable bounds, we first reorder the hypotheses for each segment in isolation, using a standard trigram model, and then perform dynamic programming (Viterbi or forward-backward) only on the top 20 hypotheses for each segment. Experiments showed that the improvements from including more hypotheses in the dynamic programming were negligible.

4.3. Results

First we compared word error performance under four different segmentation conditions and scoring algorithms:

(1) matched acoustic segmentation in training and testing, using the standard rescoring algorithm;

(2) mismatched segmentations, using the standard algorithm (language model trained on linguistic segments, rescoring on acoustic segments);

(3) mismatched segmentations, using the hidden linguistic-segment language model with the Viterbi rescoring algorithm.

(4) mismatched segmentations, using the hidden linguistic-segment language model with the forward-backward rescoring algorithm.

Results are shown in Table 1.

Table 1. Word error result for three different segmentation conditions/rescoring methods

	Model/Rescoring method	WER
(1)	Acoustic trigram/standard	53.7
(2)	Linguistic trigram/standard	54.4
(3)	Linguistic trigram/Viterbi	53.8
(4)	Linguistic trigram/Forward-backward	53.8

The absolute word error rate (WER) differences are small; yet a Sign test on the utterance-level word errors reveals a significant difference between conditions (1) and (2) ($p < 0.0001$), and between (2) and (3) ($p < 0.0005$), though not between (1) and (3).

Experiment (2) confirms that a segmentation mismatch does indeed lead to degraded language model performance. Experiment (3) shows that the hidden segmentation model can effectively compensate for this mismatch, yielding results that are close to those of the matched-segmentation language model. The comparison of Viterbi and forward-backward rescoring (Experiment 4) suggests that there is no practical advantage in using the latter, even though it is theoretically superior in optimizing utterance error.

4.4. Modeling Turn Boundaries

When comparing acoustic and linguistic segment language models one has to bear in mind that acoustic segmentations are based on turn boundaries and pauses, which constitute potentially valuable information for modeling the distribution of words. For example, certain words, such as backchannel responses ("Yeah", "Uh-huh") are more likely after turn boundaries. The linguistic segment model does not have access to these cues, and is therefore at an inherent disadvantage. This reasoning is consistent with results showing that explicit language modeling of pauses in Switchboard improves recognition accuracy [16].

The obvious solution to this problem is to include non-lexical cues such as turn boundaries and pauses in the linguistic segment language model. Since the N-best lists available to us did not contain pauses we decided to model turn boundaries only, which can be inferred from the segment boundary times. The linguistic segment language model was rebuilt from transcripts containing turn boundary tags `<t>`, i.e., the `<t>` tag was treated as a regular word. Both the original and the turn boundary model were compared using the Viterbi rescoring method; results are shown in Table 2.

Table 2. Word error results with and without turn boundary modeling

Model	WER
Linguistic trigram	53.8
Linguistic trigram with turns	53.2

As expected, explicit turn boundary modeling yields the better language model; the reduction in WER is small

(0.6% absolute), but again highly significant ($p < 0.0001$). Note that it is not meaningful to compare language model perplexities, as the two models differ in the number of both word types and tokens.

5. CONCLUSIONS AND FUTURE WORK

We showed that a hidden segmentation model can be used effectively in N-best list rescoring to overcome the mismatch between acoustic and linguistic segmentations in language modeling. The algorithm used is based on Viterbi or forward-backward rescoring of conversation-length recognition hypotheses, i.e., across acoustic segment boundaries. In particular, we showed that this method allows a language model trained on linguistic segments to achieve the same performance on acoustically segmented test data as a language model trained on matched acoustic segments. Furthermore, the linguistic segment model was improved further by modeling turn boundaries explicitly.

The work to date suggests a number of promising continuations. We expect further improvements by incorporating other non-lexical events, such as pauses, into the linguistic segment language model. We are also investigating a conversation-level language model that considers the turns of both speakers (rather than treating each conversation side as a monologue). This should capture frequent utterance/response pairs, as well as collaborative completions, where one speaker finishes the other's utterance. Finally, language modeling of segment and turn boundaries should be accompanied by an explicit modeling of the prosodic features of such events. We have started work on the combined prosodic and language modeling of hidden events [11], which we plan to extended for the purpose of N-best rescoring.

REFERENCES

[1] V. Digalakis and H. Murveit. GENONES: An algorithm for optimizing the degree of tying in a large vocabulary hidden Markov model based speech recognizer. In *Proceedings of the IEEE Conference on Acoustics, Speech and Signal Processing*, Adelaide, Australia, 1994.

[2] J. J. Godfrey, E. C. Holliman, and J. McDaniel. SWITCHBOARD: Telephone speech corpus for research and development. In *Proceedings of the IEEE Conference on Acoustics, Speech and Signal Processing*, vol. I, pp. 517–520, San Francisco, 1992.

[3] S. M. Katz. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 35(3):400–401, 1987.

[4] M. Meteer et al. Dysfluency annotation stylebook for the Switchboard corpus. Linguistic Data Consortium. Revised June 1995 by Ann Taylor.

[5] M. Meteer and R. Iyer. Modeling conversational speech for speech recognition. In E. Brill and K. Church, editors, *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 33–47, University of Pennsylvania, Philadelphia, 1996.

[6] H. Murveit, J. Butzberger, V. Digalakis, and M. Weintraub. Large-vocabulary dictation using SRI's DECIPHER speech recognition system: Progressive search techniques. In *Proceedings of the IEEE Conference on Acoustics, Speech and Signal Processing*, vol. II, pp. 319–322, Minneapolis, 1993.

[7] M. Ostendorf, B. Byrne, M. Bacchiani, M. Finke, A. Gunawardana, K. Ross, S. Roweis, E. Shriberg, D. Talkin, A. Waibel, B. Wheatley, and T. Zeppenfeld. Model systematic variations in pronunciation via a language-dependent hidden speaking mode. Research Note 24, Center for Language and Speech Processing, Johns Hopkins University, Baltimore, 1997.

[8] M. Ostendorf, A. Kannan, S. Austin, O. Kimball, R. Schwartz, and J. R. Rohlicek. Integration of diverse recognition methodologies through reevaluation of N-best sentence hypotheses. In *Proceedings DARPA Speech and Natural Language Processing Workshop*, pp. 83–87, Pacific Grove, CA, 1991. Defense Advanced Research Projects Agency, Information Science and Technology Office.

[9] L. R. Rabiner and B. H. Juang. An introduction to hidden Markov models. *IEEE ASSP Magazine*, 3(1):4–16, 1986.

[10] R. Rosenfeld, R. Agarwal, B. Byrne, R. Iyer, M. Liberman, E. Shriberg, J. Unverfuehrt, D. Vergyri, and E. Vidal. LM95 Project Report: Language modeling of spontaneous speech. Research Note 1, Center for Language and Speech Processing, Johns Hopkins University, Baltimore, 1996.

[11] E. Shriberg, R. Bates, and A. Stolcke. A prosody-only decision-tree model for disfluency detection. In *Proceedings 5th European Conference on Speech Communication and Technology*, Rhodes, Greece, 1997.

[12] A. Stolcke, C. Chelba, D. Engle, V. Jimenez, L. Mangu, H. Printz, E. Ristad, R. Rosenfeld, D. Wu, F. Jelinek, and S. Khudanpur. Dependency language modeling. Research Note 24, Center for Language and Speech Processing, Johns Hopkins University, Baltimore, 1997.

[13] A. Stolcke, Y. Konig, and M. Weintraub. Explicit word error minimization in N-best list rescoring. In *Proceedings 5th European Conference on Speech Communication and Technology*, Rhodes, Greece, 1997.

[14] A. Stolcke and E. Shriberg. Automatic linguistic segmentation of conversational speech. In *Proceedings of the International Conference on Spoken Language Processing*, vol. 2, pp. 1005–1008, Philadelphia, 1996.

[15] A. Stolcke and E. Shriberg. Statistical language modeling for speech disfluencies. In *Proceedings of the IEEE Conference on Acoustics, Speech and Signal Processing*, vol. 1, pp. 405–408, Atlanta, 1996.

[16] T. Zeppenfeld, M. Finke, K. Ries, M. Westphal, and A. Waibel. Recognition of conversational telephone speech using the janus speech engine. In *Proceedings of the IEEE Conference on Acoustics, Speech and Signal Processing*, vol. 3, pp. 1815–1818, Munich, 1987.