PROSODIC KNOWLEDGE SOURCES FOR AUTOMATIC SPEECH RECOGNITION

Dimitra Vergyri *Andreas Stolcke* *Venkata R. R. Gadde* *Luciana Ferrer* *Elizabeth Shriberg*

Speech Technology and Research Laboratory
SRI International
Menlo Park, CA, U.S.A.

ABSTRACT

In this work, different prosodic knowledge sources are integrated into a state-of-the-art large vocabulary speech recognition system. Prosody manifests itself on different levels in the speech signal: within the words as a change in phone durations and pitch, inbetween the words as a variation in the pause length, and beyond the words, correlating with higher linguistic structures and nonlexical phenomena. We investigate three models, each exploiting a different level of prosodic information, in rescoring N-best hypotheses according to how well recognized words correspond to prosodic features of the utterance. Experiments on the Switchboard corpus show word accuracy improvements with each prosodic knowledge source. A further improvement is observed with the combination of all models, demonstrating that they each capture somewhat different prosodic characteristics of the speech signal.

1. INTRODUCTION

One source of information that is currently not explicitly modeled in most state-of-the-art automatic speech recognition (ASR) systems is prosody: phone duration, suprasegmental duration, pause, pitch, and energy measurements. Prosodic features, unlike traditional segmental features (i.e. MFCC), are robust to noise and unaffected by channel conditions. Hence, modeling their interaction with words is important for improving ASR.

Prosody manifests itself on different levels in the speech signal. Within the words, phone durations and pitch depend on lexical stress and surrounding word and pause context. For example, pauses affect the vowel durations of preceding words, an effect known as "prepausal lengthening". Between the words, the variation of the pauses is another prosodic characteristic of speech. Finally, pitch, energy, pause, and vowel lengthening are correlated with higher linguistic structures and nonword phenomena, such as sentence boundaries, disfluencies, syntax, and semantics.

Past research in modeling prosody has dealt independently with the different aspects of the prosodic information. There have been studies that used prosody to help in syntactic disambiguation and understanding [1, 2] or to detect disfluencies and sentence boundaries [3, 4]. Other efforts studied the effects of lexical stress, phone durations, or higher-level prosodic information in ASR. Examples of this work can be found in [5, 6, 7, 8].

In this paper we investigate two of the above-mentioned techniques plus a novel one, in order to integrate the different levels of prosodic information in an ASR system:

- The first approach is an improved version of the word-duration model described in [6]. Each word is represented by a duration feature that is a vector comprised of the durations of the individual phones in the word. Gaussian Mixture Models (GMMs) are used to predict the word duration features from the hypothesized words and pauses.
- The second approach addresses the interaction between words and the between-word pauses, using an N-gram model to predict the duration of the pauses from the word context.
- In the third approach [8] prosodic features are used to predict certain hidden events in speech, such as segment boundaries and disfluencies. The interaction of the hidden events with the spoken words is modeled by an N-gram language model (LM).

All the above techniques were integrated in SRI's 2002 Switchboard system [9], showing a consistent improvement over the baseline. Each approach models different aspects of speech prosody, although there is some overlap between models. While, as expected, the improvements were not entirely additive, we obtained best performance when all three techniques were integrated into a single system.

2. MODEL DESCRIPTION

2.1. Word Duration Models

Traditional acoustic models computing the probability $P(O|W)$ of the acoustic observations O given a word sequence W, do not model explicitly the durational characteristics of the hypothesized words.

In this work we revisit the word-duration models described in [6]. Each word is represented by a duration feature that is a vector comprising the durations of the individual phones in the word. For example, the word "that", represented as the phone sequence "dh+ae+t", may be represented by a duration feature (10.0 8.0 4.0), where the three values represent the durations of the three phones "dh", "ae", and "t", respectively. Thus, the feature vector captures the durations of the phones within the context of the given word. Given sufficient instances of a word, we can train statistical models to represent word duration patterns. In our experiments, we used GMMs.

The duration models can be used to rescore the recognition hypotheses in an N-best list. This way, the probability of the acoustic observations can be broken into the probability of the standard acoustic features O_A and the probability of the word-duration features O_D:

$$P(O|W) = P(O_A, O_D|W) \approx P(O_A|W)P(O_D|W) \quad (1)$$

This material is based upon work supported by the Defense Advanced Research Projects Agency Information Awareness Office EARS program. ARPA Order No. N614, Program Code No. 2E20, issued by DARPA/CMO under Contract No. MDA972-02-C-0038. Approved for public release; distribution is unlimited.

assuming conditional independence given the hypothesis W.

In developing the duration models we must deal with the problem of sparse words. We are limited to the words in the data with a minimum number of occurrences. Also, since word durations are sensitive to the pauses that may follow (prepausal lengthening), we actually want to train separate models for the words followed by pauses. This makes the sparseness of data an even bigger problem. To deal with this, we train duration models of individual triphones and phones along with those of words. We apply a simple back-off scheme, in which the triphone models are used to score an unseen word. If a triphone model does not exist, we back off to the corresponding context-independent phone model.

Another problem is the rate of speech (ROS) variation across different speakers. We estimated ROS as the average number of phones per unit time, and used it to normalize the durations of the phones in both training and testing. We found that such a normalization performed at the speaker level gave the best results. We excluded nonspeech (pause and noise) phones from duration scoring, both because this gave better results in isolation and because it makes the model more orthogonal to our other prosodic models.

2.2. Pause Language Model

Standard acoustic models provide the probability $P(O|W,S)$ of the observations given the word sequence W *and* the hypothesized nonspeech regions S. Nevertheless, standard *language* models do not include pauses, which constitute an important component of prosody. However, simply including pauses in N-gram LMs would fragment the N-gram space. It would also limit the training data from which we can estimate N-gram probabilities since we must use speech, rather than text, data.

One solution is to model the pauses as a separate knowledge source, using a trigram model conditioned on the surrounding words. This way, we approximate

$$P(W,S) = P(W)P(S|W) \approx \prod_{i=1}^{N} P(w_i|h_{i-1})P(s_i|w_i, w_{i+1}) \quad (2)$$

where s_i is the length of the pause following the ith word. In our experiments we quantized pause lengths into a few discrete bins in order to obtain reliable estimates. We found preceding words to be slightly better predictors of pauses than following words; consequently, the pause LM backs off to $P(s_i|w_i)$ in the absence of a trigram.

2.3. Modeling the Prosody of Hidden Events

The models we have described capture some of the interactions between words and durations in phones and pauses. But prosody correlates also with linguistic structures beyond the words themselves, and includes cues other than durations. Taking the approach described in [8] we try to leverage prosody for word recognition by modeling certain higher-level phenomena that manifest themselves prosodically, such as sentence boundaries and speech disfluencies. We refer to these phenomena as *hidden events*, because they can be thought of as hidden pseudo-words occurring between the observable words.

If we denote by E the hidden event representations embedded in W, and by F the prosodic features associated with those events, then we want to have a model for the relation between words, hidden events, and prosody: $P(W,E,F)$. The motivation for modeling E is that it may be easier to compute the above model than $P(W|F)$ directly. We can compute the latter as

$$P(W|F) = P(W,F)/P(F) = \sum_E P(W,E,F)/P(F) \quad (3)$$

and then decompose $P(W,E,F)$ as

$$P(W,E,F) = P(W,E)P(F|W,E) = P(W,E)\prod_{i=1}^{n} P(F_i|E_i,W) \quad (4)$$

where we assume that the prosodic feature vector F_i correlates only with the event E_i. We compute the feature from a window around the boundary of that event, so this is a reasonable assumption to make.

For the *event LM* $P(W,E)$ we use standard N-gram modeling techniques on a text corpus in which the events are marked as tags following the words: $W_1E_1W_2E_2...W_nE_n$. During testing, the events are unknown, and according to equation (3) we need to sum over all possible event sequences. The joint model $P(W,E,F)$ thus becomes equivalent to an HMM, whose states are the (word,event) pairs, while the prosodic features form the observations. Transition probabilities are given by the event N-gram model; emission probabilities are given by $P(F_i|E_i,W)$. Since the event space is discrete and small, and the prosodic feature space continuous, high-dimensional and highly correlated, we invert the problem and model posterior probabilities instead:

$$P(F_i|E_i,W) \approx P(F_i|E_i) = \frac{P(F_i)}{P(E_i)}P(E_i|F_i) \quad (5)$$

where we assume that the prosodic features are marginally independent of the word sequence given the events. This is justified since for the computation of the features we use only the segmentation information associated with a word sequence and ignore the word identities. The posterior probabilities $P(E_i|F_i)$ can be estimated by a variety of probabilistic classifiers such as decision trees, neural networks, or exponential models. By resampling the classifier training data we obtain equal priors for all events, such that $P(F_i)/P(E_i)$ in (5) can be treated as constant for a given F_i.

3. MODEL INTEGRATION

We can revise the standard equation of maximum a posteriori probability (MAP) decoding to include the conditioning on the prosodic features we are using. So the MAP hypothesis WS^*, where S is the pause sequence accompanying the words, given the acoustic features O_A, the word duration features O_D, and the other prosodic features F, would be

$$WS^* = \underset{WS}{\operatorname{argmax}}\ P(W,S|O_A,O_D,F) \approx$$

$$\approx \underset{WS}{\operatorname{argmax}} \frac{P(W,S|F)P(O_A|W,S,F)P(O_D|W,S,F)}{P(O_A|F)P(O_D|F)} \quad (6)$$

$$\approx \underset{WS}{\operatorname{argmax}}\ P(W,S|F)P(O_A|W,S)P(O_D|W,S) \quad (7)$$

$$\approx \underset{WS}{\operatorname{argmax}}\ P(W|F)P(S|W)P(O_A|W,S)P(O_D|W,S) \quad (8)$$

$$\approx \underset{WS}{\operatorname{argmax}} \sum_E P(W,E,F)P(S|W)P(O_A|W,S)P(O_D|W,S) \quad (9)$$

$$\approx \underset{WS}{\operatorname{argmax}} \left(\sum_E P(W,E)P(F|E) \right) P(S|W)P(O_A|W,S)P(O_D|W,S) \quad (10)$$

In (6) we make the same conditional independence assumption as in (1). Equation (7) relies on the approximation that the observations O_A and O_D are independent of the other prosodic features F, conditioning on the word sequence. We ignore the denominator in (6) since we assume that it is constant with respect to WS. [1] In (8) we introduce the pause model as was done in (2). Finally, to obtain (9) and (10) we apply equations (3) and (4).

The final term we need to compute in (10) is a product (or alternatively a log-linear combination) of four separate knowledge sources: the prosodic hidden event model (HE), the pause LM (Pau), the standard acoustic model (AC), and the duration model (Dur). In most state-of-the-art systems we would also include a term for the pronunciation probability of the words, and one for the word insertion penalty. This being a log-linear combination of knowledge sources [11] we combine them using discriminatively optimized weights. So in practice we introduce one exponent for each of the above-mentioned knowledge sources, which is optimized to minimize the word errors on held-out data. We also introduce a separate exponent for the prosodic model $P(F|E)$, within the HE model, which is optimized separately and reflects the relative importance of the event classifier $P(E|F)$ (in equation (5)) relative to the event LM.

4. EXPERIMENTS

We tested our models on Switchboard (SWB) data from recent NIST Hub-5 benchmarks. The 2001 development set was used as the held-out set on which the exponents of the knowledge sources were optimized. We used a simplex downhill method for the optimization of the log-linear weights on an N-best hypothesis list as in [12][2].

The data from the 2001 and 2002 evaluations was used for testing. The results of our experiments are in Tables 1 and 2. Table 1 gives the MAP decoding word error rate (WER) for different systems that include each of the prosodic models separately or in combination with others. Table 2 gives results after performing an N-best ROVER combination with two other systems that used different front ends, as used in the full evaluation system [9]. As an expedient, these additional systems were not rescored with the HE model, but did use the duration model and pause LM, for the experiment that included prosodic information.

The baseline and prosodic models used in the experiments are described in more detail below:

Baseline: The baseline system (BS) used MMIE trained cross-word acoustic models (∼150K Gaussians) of PLP features with SAT and MLLR adaptation, a word-class 4-gram LM, and a pronunciation probability model. An important detail is that N-best decoding in our system used acoustic models that had been adapted using prior recognition outputs, which in turn had already been rescored once with the duration model and pause LMs. This means that the acoustic models had already benefited from some of the prosodic knowledge sources, making further improvements harder to achieve.

Pause model: The pause model (Pau) was a standard back-off-trigram model predicting only three levels of pause duration: pause<0.06sec, 0.06<pause<0.6sec, and pause> 0.6sec. Using a baseline acoustic model we performed a time alignment of all the acoustic training data utterances against their transcriptions. From these alignments, we obtained the durations of phones and pauses that were used to train the model.

Table 1. Word Error rates of the MAP hypothesis using rescoring of N-best hypotheses obtained with a PLP baseline system.

System	dev'01	eval'01	eval'02
Baseline system (BS)	28.2	26.7	29.1
BS + Pau	28.1	26.7	28.9
BS + Dur	27.6	26.4	28.6
BS + HE	27.6	26.3	28.6
BS + Pau + Dur	27.4	26.4	28.5
BS + Pau + HE	27.5	26.3	28.5
BS + Dur + HE	27.1	26.1	28.3
BS + Pau +Dur + HE	27.1	26.0	28.2

Duration model: The time-aligned acoustic training data were used to train the Duration models (Dur) for each word with a minimum of 20 training occurrences (∼8K word-models + ∼10K models which included following pause/no-pause information). Triphone and phone duration models were also trained and were used as a back-off for the rest of the words. The covariances were modeled by a full-diagonal matrix. ROS normalization was applied on the speaker level, estimating two ROS normalization parameters per speaker, corresponding to vowels and consonants.

Hidden event model: The prosodic model in (5) was trained using 900 SWB conversations annotated with hidden events by LDC [13]. A CART-type decision tree was used as our modeling approach. A vast range of prosodic features was explored, but the best tree made use mainly of the duration of the current and previous pauses, durations of last observed syllable rhymes, vowels and last stressed vowel (normalized by phone and speaker-specific statistics), the distance from last speaker turn, and a flag indicating whether or not the current boundary corresponds to a speaker turn. Less frequent, but still present in the best tree, were questions about the pitch pattern and fluctuations in the last word.

We used the tree to predict five hidden event types: sentence boundaries (occurring with a frequency of 10.8% in the training data), filled pauses (2.9%), repetitions (1.9%), deletions (1.3%), and all others (82.9%). The tree resulted in a prediction accuracy of 61.8% (compared to chance at 20%), corresponding to an entropy reduction of 42.5%.

For the $P(W, E)$ term of the Hidden Event (HE) model we trained a word-class 4-gram LM. The same data and the same word classes as in the baseline LM were used. First, the LDC event-annotated conversations were used to obtain an initial model, which was then used in a tagger to automatically annotate the rest of the SWB language model training corpus. The whole corpus was eventually used to retrain the 4-gram hidden event class LM.

5. DISCUSSION

As shown in Table 1, all prosodic models improve the baseline system to various degrees. The smallest WER reduction, 0.1-0.2% absolute, comes from the pause LM. Improvements due to the duration model range from 0.3 to 0.6% absolute over the baseline system. The hidden event model reduces the WER about as much as the duration model, and when combined with the other

[1] This is a loose approximation since the features are computed using the segmentation information associated with W. The work in [10] proposes a solution to this problem, but for this study we take the independence assumption.

[2] The algorithm is implemented in the SRILM Toolkit, available from `http://www.speech.sri.com/projects/srilm/`.

Table 2. WER of ROVER combination of 3 systems: PLP, MFCC and LFC (Fourier), with and without the use of prosodic information.

3 system ROVER	dev'01	eval'01	eval'02
no prosodic information	27.1	25.8	27.9
with prosodic information	26.2	25.2	27.2

two prosodic knowledge sources still gives an additional 0.3-0.4% improvement. All WER reductions relative to the baseline on the combined eval'01 and '02 utterances are significant in a matched pairs Sign test, $p < 0.0001$ one-sided.

Overall, the combined use of all three prosodic models lowers WER by 0.7-0.9% on the independent test sets. Even in the ROVER-combination with the two other systems that did not make use of the HE model, an improvement of 0.6-0.7% over the system combination without prosody is preserved. The gains due to each individual prosodic model were somewhat complementary, but not fully additive, as expected. This confirms that the information sources modeled are in fact correlated, and suggests that relaxing some of the independence assumptions may result in further improvements.

As already mentioned, the duration model and pause LM had been used earlier in the processing to generate adaptation hypotheses for the baseline system, thereby making incremental gains from prosody scoring harder to achieve. In fact, prior to transcription mode adaptation, compared to a baseline WER of about 32%, the duration model achieves a WER reduction of about 1% absolute, and the pause LM an additional 0.4%. Even in relative terms, these improvements are larger than those obtained after adaptation, consistent with the acoustic model having absorbed some of the prosodic information via adaptation.[3]

6. CONCLUSIONS AND FUTURE WORK

We found that the use of different levels of prosodic information in otherwise competitive large-vocabulary speech recognition systems is effective in reducing word errors. We observed improvements in WER of 0.7-1.1% absolute using MAP decoding over a PLP baseline SWB system. Using ROVER with multiple front-end systems and more knowledge sources, we found the improvement due to the prosodic information to be 0.7-0.9%.

In future work we will revisit some of the independence assumptions made in our models, and will try to make use of more prosodic features. Currently, we ignore the fact that the prosodic features computed are dependent on the current hypothesis, which makes the likelihood scores of different hypotheses not entirely comparable. In [10] a normalization "anti-phone" model is proposed to account for this error in probability estimation. Also we plan to explore more of the interaction between prosodic features like pitch and energy with the word sequence. In the current work such features were modeled only indirectly through their interaction with hidden events. Since non-default events (sentence boundaries and disfluencies) occur at only 17% of the word boundaries one would expect better results with an approach that explicitly characterizes the pitch and energy features of all words.

[3]For logistical reasons we did not apply the hidden event model to the pre-adaptation baseline, and rescored only one of the adapted systems.

7. ACKNOWLEDGMENTS

This work was partially funded by NSF STIMULATE grant IRI-9619921 and by DARPA under contract MDA972-02-C-0038. The views herein are those of the authors and do not reflect the policies of the funding agencies.

8. REFERENCES

[1] M. Ostendorf, C. Wightman, and N. Veilleux, "Parse scoring with prosodic information: an analysis/synthesis approach," *Computer Speech and Language*, vol. 4, pp. 193–210, 1993.

[2] R. Kompe, *Prosody in speech understanding systems*, Springer, Berlin, 1997.

[3] A. Stolcke, E. Shriberg, R. Bates, M. Ostendorf, D. Hakkani, M. Plauché, G. Tür, and Y. Lu, "Automatic detection of sentence boundaries and disfluencies based on recognized words," in *Proc. ICSLP*, R. H. Mannell and J. Robert-Ribes, Eds., Sydney, Dec. 1998, vol. 5, pp. 2247–2250, Australian Speech Science and Technology Association.

[4] E. Shriberg, A. Stolcke, D. Hakkani-Tür, and G. Tür, "Prosody-based automatic segmentation of speech into sentences and topics," *Speech Communication*, vol. 32, no. 1-2, pp. 127–154, Sept. 2000, Special Issue on Accessing Information in Spoken Audio.

[5] J. L. Hieronymous, D. McKelvie, and F. R. McInnes, "Use of acoustic sentence-level and lexical stress in HSMM speech recognition," in *Proc. ICASSP*, San Francisco, Mar. 1992, vol. 1, pp. 225–227.

[6] V. R. R. Gadde, "Modeling word durations," in *Proc. ICSLP*, B. Yuan, T. Huang, and X. Tang, Eds., Beijing, Oct. 2000, vol. 1, pp. 601–604, China Military Friendship Publish.

[7] N. M. Veilleux and M. Ostendorf, "Prosody/parse scoring and its applications in ATIS," in *Proc. ARPA HLT Workshop*, Plainsboro, NJ, Mar. 1993, pp. 335–340.

[8] A. Stolcke, E. Shriberg, D. Hakkani-Tür, and G. Tür, "Modeling the prosody of hidden events for improved word recognition," in *Proc. EUROSPEECH*, Budapest, Sept. 1999, vol. 1, pp. 307–310.

[9] A. Stolcke, R. Gadde, A. Venkataraman, D. Vergyri, J. Zheng, and C. Wooters, "The SRI RT-02 speech-to-text system," in *NIST Rich Transcription Workshop*, Vienna, VA, May 2002.

[10] J. Glass, J. Chang, and M. McCandless, "A probabilistic framework for feature-based speech recognition," in *Proc. ICSLP*, H. T. Bunnell and W. Idsardi, Eds., Philadelphia, Oct. 1996, vol. 4, pp. 2277–2280.

[11] 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 *Proc. DARPA SNP Workshop*, Pacific Grove, CA, Feb. 1991, pp. 83–87, Defense Advanced Research Projects Agency, Information Science and Technology Office.

[12] D. Vergyri, "Use of word level side information to improve speech recognition," in *Proc. ICASSP*, Istanbul, June 2000, vol. 3, pp. 1823–1826.

[13] M. Meteer, A. Taylor, R. MacIntyre, and R. Iyer, "Dysfluency annotation stylebook for the Switchboard corpus," Distributed by LDC, `ftp://ftp.cis.upenn.edu/pub/treebank-/swbd/doc/DFL-book.ps`, Feb. 1995, Revised June 1995 by Ann Taylor.