Prosodic Features for a Maximum Entropy Language Model

Oscar Chan, Roberto Togneri

School of Electrical, Electronic and Computer Engineering
University of Western Australia, Australia
{oscar,roberto}@ee.uwa.edu.au

Abstract

This paper presents an approach for incorporating prosodic knowledge into the language modelling component of a speech recogniser. We formulate features for a maximum entropy language model which capture various aspects of the relationships between prosody, syntax and the spoken word sequence. Maximum entropy is a powerful modelling technique, and well suited to modelling prosodic information. Tests conducted on the Boston University Radio Speech Corpus using this model showed improvements in perplexity, and n-best rescoring results also demonstrated small but statistically significant gains.

Index Terms: language modelling, prosody, speech recognition.

1. Introduction

Prosody typically refers to the suprasegmental characteristics of speech, such as intonation, rhythm and stress, which are used to communicate structural information about an utterance. The domain of prosodic features is quite broad, and ranges from the syllable up to whole utterances and discourses. In the speech signal, the acoustic properties considered to be the dominant correlates of prosody are fundamental frequency, energy and duration.

There has been an increasing interest in the use of prosody as a knowledge source for spoken language processing systems, however, it is still an area largely confined to research. One reason for this is that the nature of prosody is still not well understood. While it is generally accepted that prosody plays an important role in human perception of speech [1], it has yet to be fully exploited in automatic speech systems. This is, in part, due to the large number of functions prosody serves. In addition to conveying linguistic information, such as syntax and semantics, prosody also contains paralinguistic effects like emotion and physiological characteristics. When dealing with automatic speech recognition, these paralinguistic features carry little to no meaningful information, but isolating relevant features is a difficult task. A recent trend in automated speech processing systems employing prosody has been to extract large sets of raw features from speech data, which can then be used directly in a variety of statistical models [2, 3]. This is a sensible approach, as it enables fast, automatic processing while allowing the model to choose appropriate features, however, it is not without it's share of problems. There are a large number of factors which can influence the prosodic manifestation of an utterance, and this variability can complicate the task of reliably extracting prosodic parameters. Combining prosodic models with existing systems provides yet another challenge to overcome. For example, the Hidden Markov Model (HMM) framework of most acoustic models operates only on local information at the frame level (typically in the order of 10ms), and there are no clear methods for incorporating suprasegmental prosodic features.

Despite the difficulties outlined above, prosody has been successfully incorporated into speech systems in a variety of ways. Some examples include topic segmentation [4], disfluency detection [5], speaker verification [6] and speech recognition [7, 8]. A common approach to employing prosody in speech recognition and understanding systems has been to model the dependence between prosody and syntax. Veilleux et al. [9] used prosodic models based on prominence and break features to score syntactic parses of hypotheses in a speech understanding system. In Stolcke et al. [7], the recognition task of finding the optimal word sequence, W^*, from standard acoustic features A, was formulated as:

$$W^* = \arg\max_W p(A|W)p(W,S)p(F|W,S) \quad (1)$$

The language model component was augmented with hidden events (sentence boundaries and disfluencies), S, which were modelled with prosodic features, F. In similar work, Chen et al. [8] used Explicit Duration HMMs to include prosodic features into the acoustic modelling process, and combined these with prosodic bigram language models to create a fully prosody dependent recogniser.

Following this approach, we propose to model relationships between prosody and syntax in our work. These dependencies are modelled in a language model, as this appears to be the best way to represent syntactic and suprasegmental information. Also, language models can be easily coupled with acoustic models through the use of n-best or lattice rescoring. We have chosen to use the maximum entropy model for this task for the following reasons:

- Maximum entropy models are a well understood modelling technique, and many efficient algorithms exist for parameter estimation.
- Maximum entropy models are very flexible in that information from multiple knowledge source with differing characteristics can all be modelled in a consistent manner.
- The maximum entropy principle is especially applicable to this work. Given the lack of a clear understanding of the relationship between prosodic features and linguistic units, it is advantageous to have a framework which makes no assumptions about the data beyond what is explicitly modelled.

The layout of the remainder of this paper is as follows. In Section 2, we describe the approach used to incorporate prosodic information into a maximum entropy language model. Section 3 details the experiments and results, and conclusions are provided in Section 4.

2. Method

This section provides a brief overview of the maximum entropy modelling technique used in this study, as well as an introduction to the corpus. We then describe the features to be used in the model.

2.1. Maximum Entropy Model

A whole sentence maximum entropy (WSME) model [10] models the probability of a sentence s as:

$$p(s) = \frac{1}{Z} p_0(s) \exp\left(\sum_i \lambda_i f_i(s)\right) \quad (2)$$

where Z is a normalisation constant, $p_0(s)$ is an initial distribution (usually based on an n-gram model) and the λ_i's are real valued parameters of the model. The features, $f_i(s)$'s, are arbitrary functions of s subject to the constraints:

$$E_p[f_i] = K_i \quad (3)$$

Typically, the values of the K_i's are chosen such that they represent the empirical expectations of the features in a training corpus. Fitting the model involves finding the set of λ_i's which minimises the Kullback-Leibler divergence between the model p and the initial distribution p_0:

$$\arg\min_{p \in \mathcal{P}} D(p_0 \| p) \quad = \quad \arg\max_{p \in \mathcal{P}} \sum_s p_0(s) \log p(s) \quad (4)$$

where $\mathcal{P}$ is the set of all models with exponential form. This can be solved either iteratively [11] or with standard constrained optimisation methods.

This form of model provides great flexibility, as the $f_i(s)$'s can be any computable property of a sentence, allowing for information from multiple knowledge sources to be incorporated in a consistent manner.

Maximum entropy models, like other maximum likelihood models, are subject to over-training. As the number of features used in the model increase, the model becomes more tightly constrained to the training sample and loses its ability to generalise. In order to alleviate this, some of the probability mass can be redistributed by *smoothing*. The most common smoothing technique for maximum entropy models is to apply a Gaussian prior over the model parameters, and perform maximum a posterior optimisations performed. This effectively changes the objective function in Equation (4) to:

$$\sum_s p_0(s) \log p(s) - \frac{1}{2\sigma^2} \sum_i {\lambda_i}^2 \quad (5)$$

where the second term penalises models which diverge from the initial distribution.

2.2. Corpus

Training a statistical model such as the WSME requires a large amount of data. The Boston University Radio Speech Corpus [12] was selected for this experiment as it is one of the most widely available corpora developed for research into prosody. The corpus contains recordings of radio news broadcasts from three female and four male speakers, all of whom were professional radio announcers. It is suggested that professional speakers use clearer and more consistent prosodic structure [12], making this data suitable for analysis and use in an automated system. All utterances are sampled at 16kHz, and a subset of the data is annotated with orthographic transcriptions, phonetic alignments, part-of-speech (POS) tags from the Penn Treebank tagset [13] and prosodic labels. The prosodic transcriptions follow the Tones and Break Indices (ToBI) [14] conventions for American English. The intonational events marked by ToBI are divided into a break index tier and a tone tier. The break index tier describes the degree of juncture between each pair of words and can take values from 0 to 6, where a break index of 4 or higher represents a full intonation phrase boundary. Two types of tones are described by the tone tier. Phrasal tones represent events associated with intonational boundaries, and pitch accents represent events associated with accented syllables. The basic tones are H and L, describing high and low pitch events in the local pitch range, respectively, and more specific tone labels are constructed from these.

H*	6678
!H*	1966
L+H*	2080
L+!H*	547
H+!H*	540
L*	509
L*+H	41
no accent	11764

Table 1: *Number of pitch accents in training corpus.*

2.3. Prosodic Features

The goal of this study is to find suitable prosodic features which can be used to assist the language modelling task. We make use of the annotated data in the corpus, in particular the pitch accents. Seven types of pitch accents are defined in the corpus. Peak accents, H*, and low accents, L*, are the two basic tone types. The L*+H label represents a low tone on an accented syllable followed by a sharp pitch rise, and L+H* represents a high tone target preceded by a pitch rise. The !H* tones are a downstep onto an accented syllable, and can be preceded by a low or high tone. The distribution of words containing these pitch accents in the training data is shown in Table 1. Due to the very low rate of occurrence of some accent types, the set of labels were collapsed into two categories only - presence and absence of an accent. We also reduced break index labels in a similar manner, defining the presence of a boundary by a break index value of at least 4. Using these simplifications, transcriptions were generated such that each word was aligned with it's POS, pitch accent and break index labels.

The approach we take to modelling prosodic information is similar to [8] in that we model the relationship between prosody and syntax to acquire more robust estimates from limited data. The following features were included into the model:

- POS features: $f(s)$ = # times the word and POS sequence $(w_{i-n}, p_{i-n}, \ldots w_i, p_i)$ occurs in s. These features model the co-occurrence of word w_i with it's assigned POS tag p_i, and effectively contain n-gram type information of POS-dependent words. In our experiments, we have used unigram and bigram features of this form.
- Break index features: $f(s)$ = # times the word and break index pair (w_i, b_i) occurs in s. These features model infor-

mation about words which are likely appear at intonational phrase boundaries.

- Accent features: $f(s)$ = # times the POS and prosody sequence $(p_{i-n}, a_{i-n}, \ldots p_i, a_i)$ occurs in s. Like the POS features above, these features model the relationship between a word's POS tag p_i and it's accent label a_i. Sequences of up to length three were modelled in this way. The reason we do not include the word identity, w_i, into these features is because the pitch accents are not a reliable estimator given little contextual information. However, increasing context length in these features also greatly increase the number of parameters, hampering the ability to robustly estimate their expectations. It should be noted, though, that even when w_i is not explicitly included, the model requires no assumptions of independence between w_i and a_i to be made.
- Duration features: $f_w(s)$ = average duration of word w in s, for all w in the training set. The final set of features we used model word durations. Accentuation is correlated with increased duration of the accented syllable, and these features aim to capture information about this effect. Figure 1 plots the average durations of accented words against the average durations of the unaccented forms of those words. It can be seen that there is indeed an overall increase in duration when a word is accented. We selected a subset of these features which show the greatest variation in duration between their accented and unaccented forms, by thresholding on the value:

$$\frac{|a_i - u_i|}{u_i} \tag{6}$$

where a_i is the duration of the accented form of word w_i, averaged over all words in the training set, and u_i is the average duration of the unaccented form of the word. Words occurring fewer than three times were excluded from this analysis. The duration plot of this subset of features is shown in Figure 2.

3. Experimental Work

Experiments were carried out on a subset of the Boston University Radio Speech Corpus comprising 399 utterances. Each utterance

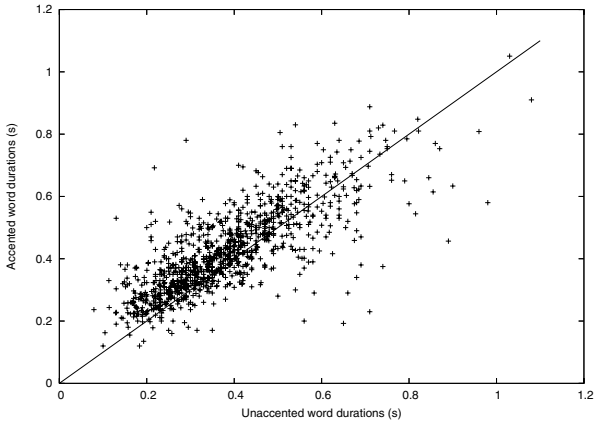

Figure 1: *Durations of accented words vs unaccented words.*

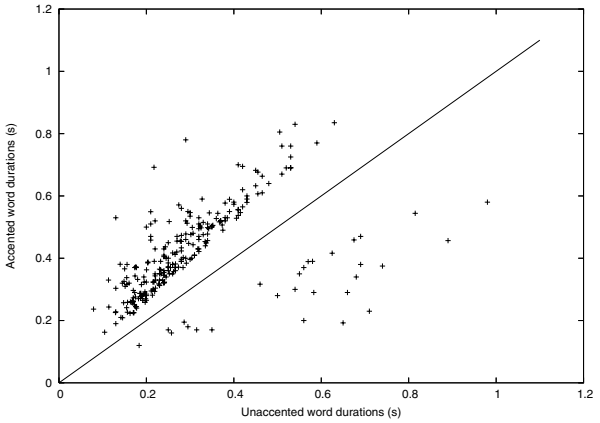

Figure 2: *Durations of thresholded words selected for the model.*

Model	PP reduction (%)	Accuracy (%)
trigram	n/a	72.73
LM-A	3.1	73.26
LM-B	6.1	73.97

Table 2: *Perplexity and n-best rescoring results of maximum entropy language models using POS features only (LM-A) and both POS and prosodic features (LM-B).*

was partitioned into individual sentences for a total 31476 words in 1529 sentences. 85% of the data were selected at random to form a training set, while the remaining 15% made up the test set. A trigram model using Good-Turing discounting was trained on the training corpus, and serves as the baseline. Two maximum entropy models were trained using the features described in Section 2.3, and smoothed with Gaussian priors. The first model, LM-A, contained only the part-of-speech features, while the second model, LM-B, used all available features. We perform both perplexity tests and word recognition tests on these models. The perplexity reduction ratio over the test set was estimated [15] as:

$$\frac{PP_{me}(T_e)}{PP_0(T_e)} = \left(\frac{Z}{\sqrt[\#s_e]{\sum_{s \in T_0} R(s)}} \right)^{\frac{\#s_e}{\#w_e}} \tag{7}$$

where $\#s_e$ is the number of sentences in the test set T_e, $\#w_e$ is the number of words in T_e and $R(s) = \sum_i \lambda_i f_i(s)$.

The two language models were also applied to the rescoring of n-best lists generated by a recogniser. The speech data was parameterised into feature vectors consisting of 13 MFCCs along with their deltas and double-deltas. These were modelled by 5 mixture Gaussians in 3-state triphone HMMs with no skips. A 100-best list was generated using this recogniser and the baseline trigram, which was then rescored using the maximum entropy models. Both perplexity and word accuracy results are presented in Table 2. LM-A, using only POS features, had a 3.1% lower perplexity than the trigram, and performing n-best rescoring improved accuracy by a relative 0.73%, with a matched pairs test [16] giving $p < 0.01$. When prosodic features were included, in LM-B, perplexity reduction over the baseline improved to 6.1%, and recognition accuracy increased to a 1.7% relative gain ($p < 0.001$).

Model	Deletions	Substitutions	Insertions
trigram	124	817	241
LM-B	131	774	223

Table 3: *Comparison of recognition errors between the trigram and LM-B.*

A more detailed breakdown of the results for LM-B is presented in Table 3. From this, it is evident that the primary source of improvements in LM-B was the reduction in substitution errors. While these gains are the result of a combination of features, we can identify some instances where specific features have contributed to the correction. An example of a corrected transcription is shown below.

trigram → ... when/WRB/u a/DT/a come/VBN/u ...
LM-B: → ... when/WRB/u they/PP/a come/VBN/u ...

In this scenario, the fragment "when they come" is correctly selected due to the presence of the feature $(p_i = \text{DT}, a_i = \text{a})$ in the original hypothesis. This feature has a very low expectation, resulting in the model giving preference to alternative hypotheses. Reductions in insertion errors were also present, however, examination of the resulting transcriptions seemed to indicate that the majority of these were side-effects of corrected substitution errors.

4. Conclusions and Future Work

In this paper, we have presented an approach for incorporating prosodic knowledge into a language model. We have argued that language models, and whole sentence maximum entropy models in particular, provide an appropriate framework for modelling suprasegmental information. Some simple features were developed for a WSME model which captured information about prosodic, syntactic and lexical dependencies. Experiments performed on the Boston University Radio Speech corpus demonstrated a 6.1% reduction in perplexity over the baseline trigram model, and n-best rescoring resulted in a 1.7% relative increase in recognition accuracy. Although the gains are small, this experiment highlights the flexibility of the WSME model. A myriad of features beyond what was considered here could plausibly be applied with little to no modification of the modelling framework. These features also need not be confined to prosody, as the only requirement for features is that they be computable properties of the sentence.

As mentioned earlier, the trend in this area of research has been to use large sets of automatically extracted features. We would also like to focus our efforts in this direction, as it opens up many avenues for exploration with regard to potential features.

5. References

[1] Alex Waibel, *Prosody and Speech Recognition*, Morgan Kaufmann Publishers, California, 1988.

[2] E. Nöth, A. Batliner, V. Warnke, J. Haas, M. Boros, J. Buckow, R. Huber, F. Gallwitz, M. Nutt, and H. Niemann, "On the use of prosody in automatic dialogue understanding," in *Proc. ESCA Workshop on Dialogue and Prosody*, Eindhoven, The Netherlands, 1999, pp. 25–34.

[3] Elizabeth Shriberg and Andreas Stolcke, "Prosody modeling for automatic speech understanding: An overview of recent research at SRI," in *Proc. ICSA Tutorial and Research Workshop on Prosody in Speech Recognition and Understanding*, New Jersey, USA, 2001, pp. 13–16.

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

[5] Christine Nakatani and Julia Hirschberg, "A speech-first model for repair detection and correction," in *Proc. 31st Annual Meeting of the Association for Computational Linguistics*, Columbus, USA, 1993, pp. 46–53.

[6] Kemal Sönmez, Elizabeth Shriberg, Larry Heck, and Mitchel Weintraub, "Modeling dynamic prosodic variation for speaker verification," in *Proc. 5th International Conference on Spoken Language Processing*, Sydney, Australia, 1998, pp. 3189–3192.

[7] Andreas Stolcke, Elizabeth Shriberg, Dilek Hakkani-Tür, and Gökhan Tür, "Modeling the prosody of hidden events for improved word recognition," in *Proc. 6th European Conference on Speech Communication and Technology*, Budapest, Hungary, 1999, pp. 307–310.

[8] Ken Chen and Mark Hasegawa-Johnson, "Improving the robustness of prosody dependent language modelling based on prosody syntax dependence," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, St. Thomas, U.S. Virgin Islands, 2003, pp. 435–440.

[9] N. M. Veilleux and M. Ostendorf, "Prosody/parse scoring and its applications in ATIS," in *Proc. ARPA HLT Workshop*, New Jersey, USA, 1993, pp. 335–340.

[10] Ronald Rosenfeld, "A whole sentence maximum entropy language model," in *Proc. IEEE Workshop on Speech Recognition and Understanding*, California, USA, 1997.

[11] Adam L. Berger, "The improved iterative scaling algorithm: A gentle introduction," Tech. Rep., Carnegie Mellon University, 1997.

[12] M. Ostendorf, P. Price, and S. Shattuck-Hufnagel, "The Boston University Radio News Corpus," Tech. Rep., Boston University, 1995.

[13] M. Marcus and B. Santorini, "Building a very large natural language corpora: The Penn Treebank," *Computational Linguistics*, vol. 19, no. 2, pp. 313–330, 1993.

[14] Kim Silverman, Mary Beckman, John Pitrelli, Mari Ostendorf, Collin Wightman, Patti Price, Janet Pierrehumbert, and Julia Hirschberg, "TOBI: A standard for labeling English prosody," in *Proc. Internaional Conference on Spoken Language Processing*, Banff, Canada, 1992, pp. 867–869.

[15] Ronald Rosenfeld, Stanley F. Chen, and Xiaojin Zhu, "Whole-sentence exponential language models: A vehicle for linguistic-statistical integration," *Computer Speech and Language*, vol. 15, no. 1, pp. 55–73, 2001.

[16] L. Gillick and S. Cox, "Some statistical issues in the comparison of speech recognition algorithms," in *Proc. International Conference on Acoustics, Speech and Signal Processing*, Glasgow, Scotland, 1989, pp. 523–535.