INTRODUCING PROSODIC CONSTRAINTS TO STOCHASTIC LANGUAGE MODELLING

Andrew Hunt

Speech Technology Research Group
Department of Electrical Engineering
University of Sydney

ABSTRACT - Recent work has shown that prosodic knowledge can be used in conjunction with syntactic analysis to improve continuous speech recognition accuracy and to implement speech understanding by resolving syntactic ambiguity. One major limitation of such work is that it is based on deterministic parsing techniques which are difficult to integrate closely with HMM-based recognition systems. This paper presents a novel way of extending conventional stochastic language modelling techniques to incorporate prosodic constraints with the aim of reducing perplexity; this produces a Stochastic Prosodic Language Model (SPLM). This approach has three major advantages over previous prosody-syntax work; training is entirely unsupervised, no prosodic or syntax analysis or labelling of training data is required, and the prosodic-syntactic constraints can be directly (and easily) employed in the Viterbi search of a speech recogniser. This paper presents the theoretical background to the SPLM arising from previous work on deterministic approaches and presents the detail of the SPLM which uses a combination conventional stochastic language modelling techniques and linear discriminant analysis.

INTRODUCTION

In recent years there has been increasing interest in the use of prosodic knowledge to improve speech recognition systems. Prosody may be applied to a number of tasks within a recognition system including the resolution of syntactic ambiguity, aiding the lexical recognition process, identifying speech errors and improving interpretation of speaker intentions. Some work has focused on the development of models which relate syntactic structure to prosodic phrasing and in turn to acoustic prosodic features (e.g. Veilleux and Ostendorf, 1993; Hunt, 1994a,b). That work aimed at using knowledge of the prosody-syntax relationship to improve speech recognition accuracy by ensuring conformance of prosodic and syntactic forms, and worked towards speech understanding by resolving syntactic ambiguity.

The syntactic framework used in the work was provided by deterministic parsers (or in some cases hand-labelled syntactic analysis). This introduced several problems. Firstly, the parsing algorithms used have not been strictly left-to-right; this makes integration with HMM-based recognisers difficult and requires additional stages in the recognition process. Secondly, because of the introduction of the additional stages, an N-best candidate approach must be adopted between the lexical recognition and prosodic processing stages; this introduces the need for a trade-off between increasing N leading to greater processing requirements and decreasing N which may lead to the elimination of the correct recognition string. Thirdly, the use of deterministic parsing techniques is itself troublesome because the parser output may include many legal (but incorrect) parses and because handling of grammatical errors is usually not robust.

This paper introduces a novel means of introducing prosodic constraints to the stochastic language modelling approach for use in continuous speech recognition. Stochastic language modelling techniques are currently used in most continuous speech recognisers because they provide good performance, can be directly utilised as a constraint in the Viterbi search of HMM-based recognisers and because, in most cases, they use unsupervised training methods. By adding prosodic constraints to these models we should reduce the perplexity of a language model. This paper presents the theoretical background for

the stochastic prosodic language model which arises from the previous work with deterministic systems and then presents the algorithm and training required to produce such a model. Finally, some initial results from the implementation of such a system are presented.

SOME RESULTS FROM THE DETERMINISTIC APPROACH

The possibility of introducing prosodic constraints to a stochastic language model arose as an extension of previous work by the author which modelled the prosody-syntax relationship using the Link Parser (Sleator and Temperley, 1991) to provide a deterministic syntactic framework. In this work, several models were developed which showed significant correspondence between prosodic phrasing and syntactic structure. These models can be divided into two categories:

- Phonological Prosody-Syntax models in which the syntactic features are used to predict a phonological representation of prosodic phrasing (break indices),
- Acoustic Prosody-Syntax models in which multivariate statistical techniques are used to model the relationship between a set of acoustic prosodic features and the syntactic features.

The relevant results from the phonological models are as follows:

- It is possible to produce an effective and significant statistical model relating prosodic phrasing (as marked by break indices; Price et. al., 1991) to syntactic structure (Hunt, 1993; 1994c).
- Further, within the link parser framework, the effect of syntax on prosodic phrasing appears to be primarily a left-to-right function (Hunt, 1994c).

The relevant results from the acoustic models are as follows:

- It is possible to automatically train several types of linear model relating acoustic prosodic features (including intensity and durations of various phonological entities) to syntactic structure (Hunt, 1994a; 1994b; and these proceedings).
- The acoustic features which are correlated with prosodic phrasing and which are related to syntactic structure occur at, or proceed, each word break; that is, the acoustic-prosodic features which reflect syntactic structure primarily operate left-to-right (Price et. al., 1991; Hunt, these proceedings).
- The acoustic prosody-syntax models can be used in automatic speech recognition to resolve syntactic ambiguity (Hunt, 1994a; 1994b).

STOCHASTIC LINK GRAMMAR

Lafferty et. al. (1992) developed a probabilistic model of the link grammar. This new model includes the n-gram family of models as a natural sub-class. Whilst the current work does not attempt to use their language modelling approach, their work did show that it is possible to capture the grammatical representation of the link parser within a stochastic model. In other words, surface word order reflects the surface syntactic structure (as captured by the link parser) and thus it is possible to build a statistical model relating the syntactic structure to surface word order.

STOCHASTIC PROSODIC LANGUAGE MODEL

The above sections present the three sources of evidence providing theoretical support for the construction of a SPLM; in summary,

- It is possible to build a statistical model between prosodic phrasing and syntactic structure (as represented by the link parser) and such a model can operate in a strictly left-to-right fashion.
- It is possible to build a statistical model which relates acoustic prosodic features to syntactic structure (as represented by the link parser) and such a model can operate in a strictly left-to-right fashion.
- It is possible to build a statistical model relating surface word order to syntactic structure, as captured by the stochastic link grammar.

Thus, it should be possible to construct a *statistical model relating acoustic prosodic features and surface word order*, a Stochastic Prosodic Language Model.

Extending the N-Gram Language Model

The n-gram formalism produces a model of the distributional characteristics of a word given the set of previous words within a sentence, as in Equation 1.

$$p(w_1, \ldots, w_n) = p(w_1) \prod_{i=2}^{n} p(w_i \mid w_1, \ldots, w_{i-1}) \qquad [1]$$

Thus, it produces probability scores in a left-to-right fashion (this arises from the successive application of Bayes' rule; Bahl et. al., 1983). The equivalent use of context in the n-gram model and in the prosody-syntax models suggests the following prosodic extension to the calculation of word sequence probabilities shown in Equation 2.

$$p(w_1, \ldots, w_n, P_1, \ldots, P_{n-1}) = p(w_1) \prod_{i=2}^{n} p(w_i \mid w_1, \ldots, w_{i-1}, P_1, \ldots, P_{i-1}) \qquad [2]$$

where P_i represents prosodic information associated with the i^{th} word break (i.e. between w_i and w_{i+1}). Note that there is a slightly different use of indices for the words and prosodic information because the number of words in an utterance is always one greater than the number of word breaks.

In a practical implementation we must limit the context length for the conditional probability. Equation 3 restricts the context to a single word providing a prosodic bigram model.

$$p(w_1, \ldots, w_n, P_1, \ldots, P_{n-1}) = p(w_1) \prod_{i=2}^{n} p(w_i \mid w_{i-1}, P_{i-1}) \qquad [3]$$

Determining the Conditional Probability

The conditional probability on the right-hand side of Equation 3 is interpreted as providing the probability of succeeding words based on the identity of the current word and the prosodic characteristics of the current word. For example, consider the prosodic characteristics of the word pairs "because of" and "because when". For the first pair there will be a tendency towards cliticisation (tight prosodic coupling) which will be reflected by shorter durations of the final segments of "because". For the second pair, there will be tendency towards a larger intervening break because of the different syntactic form; this larger break will be reflected acoustically by longer segments in "because" and possibly by a pause.

The inclusion of the prosodic features in the conditional probability will provide different estimates of word pair probabilities for different prosodic productions of each word. A number of methods for determining the conditional probability have been considered. This paper includes description and initial testing of a method which linearises and quantises the prosodic characteristics of words developed as a modification of previous work by the author (Hunt, 1994b). That work showed that Linear Discriminant Analysis (LDA) could be used to differentiate discrete syntactic forms when trained on acoustic prosodic features. This approach was modified slightly for use in the SPLM; LDA is instead used to differentiate word pairs on the basis of the acoustic prosodic features.

What LDA provides is a linear weighting for each acoustic feature so that the examples of classes are maximally separated; in this work we produce an optimal separation of words. Figure 1 shows a representation of the separation of four words following "because" based on the prosodic characteristics of "because".

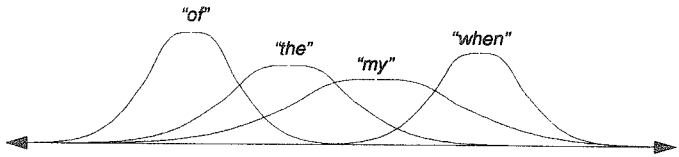

Figure 1: Separation of Succeeding Words by Linear Discriminant Analysis

For each preceding word (w_{i-1}) we produce the LDA parameters. This requires a reasonable number of examples to obtain stable estimates; where there are not sufficient examples, we can back-off to an n-gram estimate. By determining different prosodic functions for each word we provide an important characteristic. The model is not sensitive to the effects of syllable structure, word stress pattern, phonemic identity and other factors which are known to condition prosodic features.

Figure 1 represents the prosodic distributions of a succeeding word (w_i) on the basis of the prosodic characteristic of its preceding word (w_{i-1}) using a standard normal distribution. In practical terms, this has too disadvantages; not all distributions are likely to be standard normal and a large number of examples will be required to make reasonable model estimates. We overcome these problems by quantising the prosodic vector into discrete regions with approximately equal numbers of examples per region. This simplifies the calculation of probabilities to counting the distribution for each word.

To ensure sufficient data to estimate the probabilities, we can cluster succeeding words with similar prosodic distributions. This reduces the number of parameters to be estimated and increases the amount of training data for each model parameter. In the terms of Equation 3, P_i is the quantised region determined from the acoustic prosodic features, c_j is the prosodic class, and we calculate the condition probability as follows:

$$p(w_i \mid w_{i-1}, P_{i-1}) = p(w_i \mid c_j)\, p(c_j \mid w_{i-1}, P_{i-1}) \qquad [4]$$

The two conditional probabilities on the right-hand side of Equation 4 can be determined by counting examples in a training corpus. This allows for the implementation of various established smoothing techniques to improve probability estimates; e.g. back-off or deleted interpolation.

RESULTS

Initial investigations have been carried out to determine the effectiveness of the SPLM. The tests have been based on the British English Wall Street Journal database - WSJCAM0 (Robinson et. al., 1994). The database is a British equivalent of the American English WSJ0 corpus. It contains 140 speakers each speaking around 110 utterances. Of the 110 utterances, 18 are adaptation sentences and the remainder are available for training or testing of the SPLM. Of the 140 speakers, 92 are training speakers, 20 are development test speakers and there are two sets of 14 evaluation test speakers. A set of phonemic labels generated by forced recognition using the Cambridge Engineering recogniser were used in this work.

The training section of the corpus provided approximately 160,000 words with a total lexicon of around 9,800 words. This is clearly not sufficient to train a robust bigram (or SPLM) model. However, it is sufficient to evaluate the application of LDA to the task.

Word, rhyme, vocalic nuclei and pause durations were automatically extracted for the training data. An LDA model was trained for each word for which at least 100 training examples were available. These models showed good separation of the succeeding words; the estimates had correlations of greater than 0.5 for over 90% of words and greater than 0.75 for around half the words (this is a measure of the discriminating capability of LDA). These results suggest that the use of LDA to differentiate succeeding words should be effective; in other words, it is possible to improve the prediction of the word sequence by taking into account prosodic information. It is not yet clear how well this discriminating capability will generalise from training to test data.

Initial tests have been carried out to evaluate the perplexity of the SPLM on the corpus and to compare it with a bigram model. These tests have proven inconclusive because of the relatively small amount of training data (in comparison to the lexicon size). Currently, the SPLM performance is slightly worse than bigram performance when evaluated according the model perplexity. These results suggest that a more appropriate corpus is required to fully evaluate the SPLM approach. A possible candidate is ATIS which has a constrained task and smaller lexicon than WSJCAM0.

DISCUSSION

The SPLM presented in this paper has several advantages over previous prosody-syntax modelling techniques. Firstly, it should be possible to efficiently implement the model within the Viterbi search of a conventional HMM-based recogniser and this may facilitate reduction in the beam width of a recogniser. Secondly, because it is possible to implement an unsupervised training method, no syntactic or prosodic labelling of the training corpus is required which will allow training of the SPLM on very large speech corpora and rapid retraining as new corpora become available. Thirdly, whilst the model in this paper is presented as an extension to the conventional bigram model, it should be possible to implement similar extensions to other stochastic modelling approaches including trigrams and class-based models. Finally, by representing prosodic information as discrete events, we can employ a range of smoothing techniques and model enhancement algorithms which have been developed in the field of stochastic language modelling.

The most significant limitation of the SPLM is that it will require large amounts of training data (at least 100,000's of words of recorded speech). Another limitation is that it will require retraining for new tasks due to differences in both the bigram probabilities and likely changes in prosodic characteristics. Finally, retraining will also be required for changes in speaking styles.

Initial tests have shown some promise for the SPLM. However, they have demonstrated the need for a large and appropriate corpus to test the model. Work is continuing with the evaluation of the SPLM with

the particular aim of testing on a more appropriate corpus than WSJCAM0, such as ATIS which has a smaller lexicon, lower perplexity and spontaneous speech.

ACKNOWLEDGEMENTS

The author is supported by a Telecom Australia Research Laboratory Fellowship and an Australian Postgraduate Research Award. Thanks to Robin King for his helpful comments on the manuscript and to Tony Robinson (Cambridge University) for helpful discussions and the WSJCAM0 labels used in testing the model.

REFERENCES

Bahl, L.R., Jelinek, F. and Mercer R.L. (1983) *A Maximum Likelihood Approach to Continuous Speech Recognition*, IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 5, pp 179-190.

Hunt, A.J. (1993) *Utilising Prosody to Perform Syntactic Disambiguation*, Proc. 3rd European Conf. on Speech Communication - Eurospeech '93, pp 1339-1342.

Hunt, A.J. (1994a) *A Generalised Model for Utilising Prosodic Information in Continuous Speech Recognition*, Proc. 1994 IEEE Intl. Conf. on Acoustics, Speech and Signal Processing, pp II-169-172.

Hunt, A.J. (1994b) *A Prosodic Recognition Module Based on Linear Discriminant Analysis*, Proc. 1994 Intl. Conf. on Spoken Language Processing, Kobe, Japan, pp 1119-1122.

Hunt, A.J. (1994c) *Improving Speech Understanding through Integration of Prosody and Syntax*, Seventh Australian Conference on Artificial Intelligence, 1994.

Hunt, A.J. (these proceedings), *Two Linear Models Relating Acoustic Prosodics and Syntax*, Fifth Australian Conference on Speech Science and Technology, Perth, 1994.

Lafferty, J., Sleator, D. and Temperley, D. (1992) *Grammatical Trigrams: A Probabilistic Model of Link Grammar*, Technical report CMU-CS-92-181, School of Computer Science, Carnegie Mellon University.

Price, P.J., Ostendorf, M., Shattuck-Hufnagel, S. and Fong, C. (1991) *The Use of Prosody in Syntactic Disambiguation*, Jnl. Acoustical Society of America, Vol. 90(6), pp 2956-2970.

Robinson, T., Fransen, J., Pye, D., Foote, J. and Renals, S (1994 - submitted manuscript) *WSJCAM0: A British English Corpus for Large Vocabulary Continuous Speech Recognition*, Cambridge University English Department.

Sleator, D. and Temperley, D. (1991) *Parsing English with a Link Grammar*, Technical Report CMU-CS-91-196, School of Computer Science, Carnegie Mellon University.

Veilleux, N.M. and Ostendorf, M. (1993) *Probabilistic Parse Scoring with Prosodic Information*, ICASSP '93.