TWO LINEAR MODELS RELATING ACOUSTIC PROSODICS AND SYNTAX

Andrew Hunt

Speech Technology Research Group
Department of Electrical Engineering
University of Sydney

ABSTRACT - This paper presents two models based on multivariate statistical techniques which show a significant linear relationship between acoustic prosodic features and syntactic structure for professionally read speech. The models differ from most previous research in three important ways; they do not use a predetermined intermediate phonological representation but instead "learn" one, the models capture a broad range of prosody-syntax effects instead of focusing on effects at major boundaries, and the Link Parser is used to provide the syntactic framework instead of constituent parsing. The models show correlations between the acoustic prosodic and syntactic domains of 0.78 and 0.84. The role of individual acoustic and syntactic features will be analysed through the use of the multivariate models.

INTRODUCTION

It is generally accepted that there is a strong link between prosodic and syntactic structures of utterances but that this relationship is not isomorphic. Work by Gee and Grosjean (1983) based on performance structures showed that it may be possible to develop algorithms which transform syntactic trees into prosodic trees. More recent work (Wightman, 1992) has shown that it is possible to use corpus-based training methods (using Classification and Regression Trees - CART) to effectively predict break indices (which reflect prosodic phrasing) from a set of syntactic features and that these predictions can be used to resolve syntactic ambiguity. (Break indices were introduced by Price et. al., 1991.) Veilleux et. al. (1992) showed also that break indices can be predicted from a set of acoustic features and can be combined with Wightman's system to provide an automatic system for resolving syntactic ambiguity. Thus, her work showed that it is possible to build models which relate acoustic features to syntactic structure through an intermediate phonological representation, and that these models can be used in speech recognition systems to resolve syntactic ambiguity and to improve recognition accuracy.

Work by Steedman (1991) and by Hunt (1993, 1994a,b) has suggested that the use of a different syntactic framework may lead to better modelling of the relationship between prosodic and syntactic structures; in other words, constituent structure analysis, which is used in most work in the field, may not necessarily provide the best predictive framework for explaining prosodic variations. Steedman's work suggests that the surface syntactic structure and intonational structure for English are the same and can be captured by a single unified grammar based on the Categorical Combinatorial Grammar. Hunt's work showed that the link grammar provides an effective framework for resolving syntactic ambiguities in speech recognition systems (this is discussed in more detail below). CCG and the link grammar share some similarity of representational form and both have some similarity to Dependency Grammars.

The work presented in this paper analyses the phonetic aspects of two prosody-syntax models which were developed for use in ASR. The two models are based on multivariate statistical techniques (and are named after those techniques); the models are the Canonical Correlation Analysis (CCA) model (Hunt, 1994a) and the Linear Discriminant Analysis (LDA) model (Hunt, 1994b). The practical issues concerning the integration of the models into speech recognition and limitations of the parser are dealt with in the previous papers.

The most important result from these models from a phonetic viewpoint is that they show a statistically significant direct linear relationship between low-level acoustic prosodic features and higher level syntactic features for professionally read speech; the correlations between the two domains for the two

models are 0.78 and 0.84. The models are also important because they are able to "learn" sensible intermediate (or phonological) representations and thus do not require any hand-labelled speech data for the study of prosody-syntax effects. However, although this work shows a strong linear relationship between acoustic prosodics and syntax, it does not exclude the possibility that there are additional non-linear effects.

SPEECH CORPUS, ACOUSTIC AND SYNTACTIC ANALYSIS

Speech Corpus

A corpus of ambiguous sentences (Price et. al., 1991) is used in the training and analysis of both models presented in this paper. This corpus was constructed to test the ability of human subjects to interpret syntactically ambiguous sentences on the basis of prosodic features alone. Various automated and semi-automated systems have been developed for the same task (e.g. Veilleux et. al., 1992; Veilleux and Ostendorf, 1993a; Hunt, 1993; Hunt, 1994a,b).

The corpus consists of pairs of sentences with the same set of phones (and in most cases the same set of words) but with different syntactic structure and meaning. There are seven types of structural ambiguity used, each with five pairs of sentences (a total of seventy sentences) and all sentences were read by four professional news-readers with disambiguating contexts. The sentences were phonetically labelled with the SRI Decipher System. The use of professional read speech will affect the prosodic content of the corpus and must be taken into account.

Link Grammar

The work presented here uses the syntactic framework provided by the Link Grammar developed by Sleator and Temperley (1991). The parser is based upon the observation that for most sentences in most languages links can be drawn between syntactically related words without those links crossing.

The output of the link parser is a link diagram which shows the topology of the syntactic links between words. Each link is labelled to indicate its type. Figure1 shows the link diagrams for the two interpretations of a sentence from the corpus of syntactically ambiguous sentences. The example sentence exhibits ambiguity of prepositional phrase attachment, indicated by the two forms of attachment of "in German". The link labels can be interpreted as follows; "S" and "O" link a subject/object to the verb, "D" links a determiner to a noun, "J" links a preposition to the noun of the PP, "M" and "EV" link a preposition to its governing noun/verb. In total, the parsing of the ambiguous sentence corpus uses 40 different link labels.

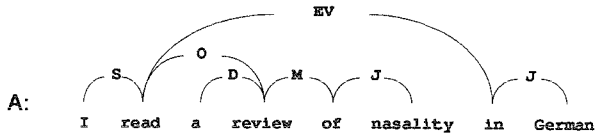
Case for the Link Grammar

The link parser was chosen for this work because of the following observations:

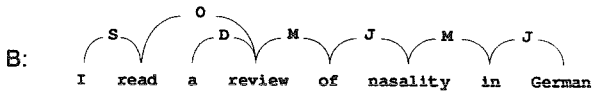
- Link diagrams show explicit *syntactic coupling* between words,
- Break indices reflect *prosodic coupling* between words,
- The similarity between these representations at an abstract level may provide a useful framework for predicting prosodic phrasing.

A pragmatic reason for selecting the link grammar framework was that a fast, efficient, public domain parser was available which covers a wide range of English syntactic phenomena.

More recent work by the author on prosody-syntax models has supported the possibility that the link grammar is a particularly effective framework for predicting prosodic phrasing. Hunt (1993) presented a semi-automatic recognition system based on the link grammar framework. This system achieved better performance than human subjects on the syntactic disambiguation task (Price et. al., 1991) and



(Far Attachment - the review is written in German)



(Near Attachment - the review concerns nasality in German)

Figure 1: Link Diagrams for an Example Ambiguous Sentence.

significantly better performance than a more complex system based on constituent analysis (Veilleux and Ostendorf, 1993a).

This work was then extended to study more subtle syntactic effects arising from topological variation in the syntactic structure using multiple linear regression (Hunt, 1994c). The most important result was that it can be shown that in the framework of the link grammar the left syntactic context has a highly significant effect upon prosodic phrasing, but that right syntactic context is not significant. This result suggests that a strictly left-to-right analysis can be effectively used for analysing prosody and syntax.

Syntactic Features

From the link diagram we extract a set of eight syntactic features representing the syntactic structure at each word break (the same set as used in Hunt 1994a,b,c). The most important of these is the link *label* which is the label of the most immediate link crossing each word pair. The remaining seven features are numeric features which provide information on the topology of the link diagram.

Acoustic Prosodic Features

A set of ten acoustic features is extracted from the speech signal. These features are either prosodic features (syllable and phoneme durations, pause duration, stress, energy and power) or are features which are known to condition prosodic features (phoneme identity, number of phonemes in a syllable, number of syllables in a word). These acoustic features are the same as were used in Hunt (1994a,b). The list of features is provided in Table 1 in the next section.

ACOUSTIC MODELS

The methodology used in the development of the CCA and LDA models was to produce an effective phonetic model relating prosodic and syntactic features and to then use the model as the base for a recognition module; a similar approach has been presented by Ostendorf et. al. (1993). This requires a model which can directly relate a set of acoustic features (to be automatically extracted from the speech signal) and a set of syntactic features (automatically extracted from the link diagrams of candidate sentences). CCA and LDA were chosen as the appropriate multivariate techniques because they could optimally model the relationship between the two domains without the need to use a predetermined intermediate representation; thus in one sense the models are unsupervised.

CANONICAL CORRELATION ANALYSIS MODEL

CCA is a multivariate statistical technique which is able to find the linear combination of vectors from two domains such that the two combinations have maximum correlation. Moreover, it is able to find successive pairs of vectors with maximum within-pair correlation but uncorrelated to the previous pairs; these pairs have decreasing correlation, hence the name "canonical" correlation. CCA was applied to the task because of the symmetric way in which it could relate the acoustic and syntactic domains.

Because the acoustic and syntactic feature sets each included a categorical parameter (*link* and *phoneme-identity*) it was necessary to develop an automatic means of assigning a numeric value to possible categorical values; the algorithm is presented in (Hunt, 1994a). The expectation was that the numeric values assigned to the *link* feature should be related to the mean index values calculated in (Hunt, 1993); these values are indeed significantly correlated ($\rho = 0.58$, $p < 0.001$). The expectation for *phoneme-identity* was that it should be related to the intrinsic duration of the vowel of the pre-boundary syllable; again these values are significantly correlated ($\rho = 0.64$, $p < 0.001$). The importance of these results is that an unsupervised model is able to learn sensible phonological values even where the input data contains no explicit representation of either break indices or intrinsic duration values.

Having produced numeric values for the categorical features we can train a full CCA model to relate prosody and syntax. The significance of the model was tested and is highly significant ($p << 0.001$). This shows that we can use CCA to produce a linear model relating low-level acoustic prosodic features to syntactic structure. Further, testing showed that at least seven CCA correlations were significant indicating that there is considerable depth to the prosody-syntax relationship. The first CCA pair has correlation of 0.78; this is a reasonably high value and along with the results already presented suggests that the linear model is a reasonably accurate and significant model of the prosody-syntax relationship. This, however, does not discount the possibility of non-linear effects.

Another hypothesis (Hunt, 1994a) was that the first intermediate value produced by the CCA vectors should be related to break indices. These values are indeed significantly correlated ($\rho = 0.62$, $p < 0.001$). This shows that a non-supervised linear model is able to learn a sensible intermediate representation which is similar to an established phonological representation of prosodic phrasing. One implication of this result is that it is possible to build effective prosody-syntax models without the need for large corpora of prosodically labelled speech.

One of the useful features of the two linear models presented here is that the role of each input feature for the model can be studied. Table 1 presents the list of acoustic prosodic features used in the CCA model and the relative importance of the feature to the model. The relative importance is defined as the weighting for the feature if it is normalised prior to the training of the CCA model (normalisation does not affect the results of CCA). The most important of the syntactic features (by a factor of 5) was the *link-label* feature; this result is in agreement with previous work (Hunt, 1993, 1994c). The relative importance of the topological features are not presented here because they are difficult to interpret without a detailed description of the features.

LINEAR DISCRIMINANT ANALYSIS MODEL

One of the limitations of the CCA model was that the syntactic features had the same weighting in all contexts. It has been shown, however, that the syntactic features take on different roles in different contexts (Hunt, 1994c) and the CCA model is unable to capture some of the more subtle effects of syntax on prosody. The LDA model was developed to address this issue.

LDA is a multivariate statistical technique which finds a linear combination of features so that samples belonging to different classes are maximally separated. Like CCA, it can produce successive linear combinations which provide decreasing separations which are uncorrelated with previous combinations (it is often referred to as Canonical Discriminant Analysis). Because the *link-label* feature had consistently been the most important syntactic feature an LDA model was trained using the acoustic values to separate the *link-label* classes. A "topological correction factor" trained with multiple linear

Acoustic Prosodic Feature	Relative Importance	
	CCA Model	LDA Model
Is any syllable in the word stressed	1.000	1.000
Phoneme identity of the vowel in the final syllable	0.404	-
Duration of the rhyme of the final syllable	0.381	0.310
Number of syllables in the word	0.295	0.527
Energy in the final vowel	0.236	-0.009
Pause duration	0.224	0.173
Intensity of the final vowel	-0.200	0.052
Duration of the final vowel	0.195	0.268
Is the last syllable stressed	0.155	0.244
Is the last syllable based on a syllabic consonant	0.135	0.137
Number phonemes in the rhyme of the final syllable	-0.119	0.072

Table 1: Relative Importance of the Acoustic Prosodic Features (ranked according to the absolute value for the CCA model)

regression was then used to reduce the model errors. (The detail of this model is presented in Hunt, 1994b).

The results and interpretation for the LDA model are very similar to the CCA model despite the use of a substantially different statistical framework. Again the first LDA discriminant vector is significantly correlated with the break index values ($\rho = 0.61$, $p < 0.001$) indicating that the model automatically learns a sensible intermediate representation despite being unsupervised. At least the first six discriminant vectors are significant indicating that there is considerable depth to the relationship. The correlation between the first discriminant values and the best estimate for the *link-label* class values is quite high ($\rho = 0.75$, $p < 0.001$) suggesting that the linear model is a reasonable way of modelling the prosody-syntax relationship. Finally, there is a high correlation between the acoustic prosodic and syntactic domains ($\rho = 0.84$, $p < 0.001$).

Table 1 presents the relative importance of the acoustic prosodic features in the LDA model. The two models are in general agreement regarding the role of the features, apart from intensity and energy. (The *phoneme-identity* feature is categorical and cannot be used with the existing LDA training method.)

DISCUSSION AND CONCLUSION

Two linear models which relate acoustic features and syntactic features have been presented which are based on partially non-supervised training methods using multivariate statistical techniques. Both models show strong and highly significant linear relationships between the acoustic prosodic and syntactic domains with correlations of 0.78 and 0.84 for the CCA and LDA models respectively. Further, both models learn phonologically sensible intermediate representations despite being unsupervised at this level. This results suggests that it may be possible to build extensive prosody-syntax models without the need for large corpora with prosodic labels.

The speech corpus used in the development of both the CCA and LDA models contains professionally read speech. It is unlikely that the models would be as accurate on non-professional or spontaneous speech; this remains an open question. However, previous work (Veilleux and Ostendorf, 1993b) has shown that the models originally developed on professional read speech can operate effectively on spontaneous speech after appropriate retraining.

The results regarding the prosody-syntax relationship and previous results discussed in the paper suggest that the link grammar provides an effective framework for analysing the relationship. Moreover, it is possible that the link grammar may provide a better framework than conventional constituent analysis (though it may not be possible to prove this either way).

Both models have been used in speech recognition tests and have provided good performance in a syntactic disambiguation task, though the performance is approximately 12% below human capability. It may also be possible to apply the models to speech synthesis applications.

There is work ongoing towards improving the acoustic representation to remove speaker-dependent and intrinsic phonemic variations and to improve the use of the syntactic features. This work has also lead to the development of prosodic extensions to stochastic language modelling techniques.

ACKNOWLEDGEMENTS

The author gratefully acknowledges Mari Ostendorf for providing the speech data used in this work, and Daniel Sleator and Davy Temperley for their assistance with the Link Parser. The author is supported by a Telecom Australia Research Laboratory Fellowship and an Australian Postgraduate Research Award.

REFERENCES

Gee, J.P. and Grosjean, F. (1983) *Performance structures: A psycholinguistic and linguistic appraisal*, Cognitive Psychology, Vol. 15, pp 411-458.

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, pp 1119-1122.

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

Ostendorf, M., Price, P.J. and Shattuck-Hufnagel, S. (1993) *Combining Statistical and Linguistic Methods for Modelling Prosody*, Proc. ESCA Workshop on Prosody, Lund, Sweden, pp 272-275.

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

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.

Steedman, M. (1991) *Structure and Intonation*, Language, Vol. 67(2), pp 260-296.

Veilleux, N.M., Ostendorf, M. and Wightman, C.W. (1992) *Parse Scoring with Prosodic Information*, Proc. 1992 Intl. Conf. on Spoken Language Processing, pp 1605-1608.

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

Veilleux, N.M., Ostendorf, M. (1993b) *Prosody/Parse Scoring and its Application in ATIS*, DARPA '93.

Wightman, C.W. (1992) *Automatic Detection of Prosodic Constituents for Parsing*, PhD Dissertation, Boston University.