Combining Statistical and Linguistic Methods for Modeling Prosody[1]

M. Ostendorf* P. J. Price† S. Shattuck-Hufnagel‡
*ECS Engineering, Boston University, Boston, MA 02215 USA
†SRI International, Menlo Park, CA 94025 USA
‡Speech Communication, RLE, MIT, Cambridge, MA 02139 USA

ABSTRACT

We describe a general approach to computational modeling of prosody that combines statistical models with linguistic theory. Statistical models provide a mechanism for representing variability, for automatically training parameters, and for analyzing large corpora. Linguistic theory provides model structure and guides feature extraction. We illustrate the approach with examples from our own work and from the work of others.

INTRODUCTION

Recently, prosody research has seen increased use of corpus-based analyses and automatic learning techniques. In particular, statistical techniques have played an important role in advancing our understanding of prosody, as well as our ability to model prosody computationally for automatic speech processing. Still, these techniques, which are driven by the need to develop robust and portable modeling techniques, are currently underutilized because of cultural differences among linguists, computer scientists and engineers. Here we try to help bridge the multi-disciplinary gap by outlining some statistical methods and providing examples of their use when driven by linguistic theory.

Statistical techniques have long been used in speech research in the analysis of significance and in socio-linguistics for sampling of large corpora, but some researchers have been wary of more extensive use because the development of linguistic insights has often been ignored in statistical modeling. However, the use of prosody in automatic speech processing cannot ignore advances in speech recognition using statistical techniques, which have the advantages that they model variability (e.g. randomness due to incomplete knowledge of sources of variability) and that automatic training methods exist for porting or adapting the models to different speaking styles or domains. Further, in the scientific aim to understand human speech communication, statistical techniques are important in that they enable the use of large corpora, which is important because human intuitions can under-represent the full range of prosodic structure that can be uncovered through the analysis of large corpora. In addition, the use of large corpora can provide data representative of normal communication, while reducing the need to control context.

Of course, statistical techniques do not provide all the answers. Stochastic models can have burdensome data requirements unless constrained by linguistic structure, and analysis of large corpora is more informative if driven by questions raised in linguistic

[1]Research was jointly funded by NSF and DARPA, NSF grants IRI-8805680 and IRI-8905249.

theory. Thus, combining linguistic and statistical methods can provide insights and results beyond the reach of either approach alone. In the next sections, we further elaborate on this theme by describing general methods for combining statistics and linguistics, illustrating these with specific examples, and discussing future directions.

COMBINING LINGUISTICS AND STATISTICS

In the context of prosody modeling, we use "linguistics" to include both phonological models of abstract units (i.e. prosodic phrase constituents, prominence, and intonation markers) and phonetic hypotheses about their observed acoustic correlates (i.e. f0, duration and energy). By "statistics", we mean both statistical data analysis and modeling techniques, recognizing two roles for statistical techniques in prosody research: (1) to generate and test hypotheses about factors that influence the phonetics and phonology of prosodic patterns, as well as to assess our level of understanding of sources of variability, and (2) to model prosodic patterns for automatic speech processing.

In **data analysis**, linguistics can provide hypotheses to test statistically, and/or we can use distributional analyses to generate hypotheses to test with traditional perception and production experiments. For variables with different interrelated conditioning factors, multi-factor analysis techniques may provide more powerful tools than significance tests. Automatic clustering and techniques for estimating model order may help answer (or pose) questions related to the number of abstract units needed to represent different prosodic phenomena. Finally, multi-modal and hierarchical models may expose distribution differences not evident from mean computations.

In **computational modeling**, linguistic theory provides the model structure, reducing dimensionality to a practical size, and drives the signal processing or feature extraction. Examples are in the next section; here, we describe general statistical techniques appropriate for prosody modeling. Classification and regression trees (or, decision trees) are particularly useful for handling a combination of categorical and continuous variables, all of which are dependent, a common situation in prosody. Decision trees take a vector of features as input, and predict or estimate a variable after a series of binary questions about the features, modeling their dependence without making explicit prior assumptions. In prosody modeling, decision tree variables draw on features traditionally used in synthesis and/or recognition rules, but the sequence and number of rules (tree structure) and their threshold values are learned automatically. While decision trees are powerful, they can only predict a single variable or vector; other techniques are needed for handling sequences of variables (random processes). The most common techniques used are the Markov source model, and the hidden Markov model (HMM), which combines a Markov chain with a random observation model. Both types of process models can incorporate decision trees to handle non-homogeneous features.

PROSODY MODELING EXAMPLES

Data Analysis. In developing a model with many variables, tabulating results for all combinations of factors is impractical. Some alternative analysis methods are illustrated in [11] for duration modeling. Automatic clustering can also be a tool for data analysis,

as in our experiments investigating questions about the categorical vs. gradient nature of acoustic differences among prominences. Decision tree design, a form of clustering, can yield insights into the relative importance of different variables, though in our work they have mainly served to confirm linguistic intuitions (e.g. hesitations are most likely to occur at a function-word/content-word boundary).

Our recent work on early accent placement within lexical items illustrates how corpus-based analysis can suggest new hypotheses. In a distributional analysis, we noticed that adjacent-stress words and alternating-stress words behaved differently with respect to within-word prominence, e.g. double accents were common only for alternating stress words. Although scattered clues might have suggested that adjacent-stress words form a special class, e.g. stress markings for these words are not consistent across dictionaries, the analysis of a large corpus made the systematic difference clearly visible.

Computational Models. As several research sites (particularly AT&T Bell Labs and ATR) have shown, corpus-based models can be powerful tools for *text-to-speech synthesis.* In prosody prediction, there are several models based on classification and regression trees, some using the predicted values associated with terminal nodes in the tree and others the probability distributions. Classification trees have proved useful for predicting abstract units [9, 8], e.g., prosodic phrase structure, pitch accent location and tone labels. Regression trees, used for estimating continuous variables, have been mainly applied to duration modeling, either for directly predicting segment duration [5, 6] or for deriving the terms in a parametric model [2].

Two aspects of prosody modeled in *speech recognition* systems are stress and duration. Several sites have used separate models for lexically stressed and unstressed vowels, though results have been mixed. Efforts in duration modeling, motivated by the fact that recognition errors often correspond to unlikely segment durations, use linguistic knowledge to define possible conditioning contexts for statistical models, e.g. [2, 5].

In *speech understanding,* where meaning of an utterance is extracted, prosody can provide information for determining the correct syntactic and semantic structure. For example, prosody has been used to reduce parsing ambiguity by automatically recognizing prosodic breaks using a decision tree and then using these breaks in a parser [4], or alternatively by scoring sentence parse hypotheses according to the likelihood of observed prosodic patterns [7]. Prosody can also provide information for speech understanding in semantic processing, since automatically detected phrasal prominence (e.g. [10, 1]) can provide clues to semantic focus. Finally, prosody can aid in detecting and correcting disfluencies such as word fragments, as in our work with decision trees.

A limitation of statistical modeling is the need for labeled training data. In speech recognition the labels are words, which can be hand-transcribed at a much lower cost than prosodic labels. Thus, it is critical to develop *automatic labeling* algorithms to assist this process. Our most successful efforts in automatic labeling have involved the use of decision trees, e.g. [10], which have outperformed HMMs.

CONCLUSIONS

Having argued the merits of combining statistics and linguistics in prosody research,

we conclude with two areas ripe for further research. First, bidirectional models are important to study, since synthesis and recognition share many of the same problems. Since the speech synthesis and recognition communities have not intersected much, they have not benefited from combining their separate perspectives. There exist a few examples using the same model for both recognition and synthesis problems, e.g. [7] for prosodic constituents, but more work is needed. Second, although it is possible to use results of statistical analysis to help formulate linguistic hypotheses, there has actually been very little such work. Two possible tools are decision trees (e.g., for ranking prediction variables), and probabilistic information measures (e.g. entropy and mutual information) that may be used to assess our ability to account for observed acoustic or phonological variability. Perhaps the biggest barrier to be overcome is the amount of knowledge required from the different disciplines, but this simply argues for multi-disciplinary collaborations. To quote Ladefoged [3]: "We all have to rely on other people to fill in the gaps – the vast holes – in our knowledge. Any scientist today is part of a team that cannot hope to build a bridge to the future without a lot of help."

REFERENCES

[1] F. Chen & M. Withgott (1992), "The Use of Emphasis to Automatically Summarize a Spoken Discourse," *Proc. Inter. Conf. on Acoust., Speech and Signal Proc.,* I229-I232.

[2] C. Fong (1993), *Duration models for speech synthesis and recognition,* M.S. Thesis, Boston University.

[3] P. Ladefoged (1992), "Knowing Enough to Analyze Spoken Languages," *Proc. Inter. Conf. on Spoken Language Proc.,* pp. 1-4.

[4] M. Ostendorf, P. Price, J. Bear, & C. Wightman (1990), "The Use of Relative Duration in Syntactic Disambiguation," *Proc. DARPA Workshop on Speech and Natural Language,* pp. 26-31.

[5] J. Pitrelli (1990), *Hierarchical Modeling of Phoneme Duration: Application to Speech Recognition,* MIT PhD thesis.

[6] M. D. Riley (1992), "Tree-based Modeling of Segmental Durations," *Talking Machines: Theories, Models and Designs,* ed. by G. Bailly, C. Benoit and T. R. Sawalis, (Elsevier Science Publishers), pp. 265-274.

[7] N. Veilleux & M. Ostendorf (1993), "Probabilistic Parse Scoring with Prosodic Information," *Proc. Inter. Conf. on Acoust., Speech and Signal Proc.,* pp. II51-54.

[8] M. Ostendorf & N. Veilleux (1993), "A Hierarchical Stochastic Model for Automatic Prediction of Prosodic Boundary Location," *J. Comp. Linguistics,* to appear.

[9] M. Wang & J. Hirschberg (1992), "Automatic classification of intonational phrase boundaries," *Computer Speech and Language,* Vol. 6, No. 2, pp. 175-196.

[10] C. Wightman & M. Ostendorf (1992), "Automatic Recognition of Intonation Features," *Proc. Inter. Conf. on Acoust., Speech and Signal Proc.,* pp. 221-224.

[11] J. van Santen and J. Olive (1990), "The analysis of contextual effects on segmental duration," *Computer Speech and Language,* Vol. 4, pp. 359-390.