Direct Modeling of Prosody: An Overview of Applications in Automatic Speech Processing

Elizabeth Shriberg *Andreas Stolcke*

Speech Technology and Research Laboratory
SRI International, Menlo Park, CA, USA
and
International Computer Science Institute
Berkeley, CA, USA

{ees;stolcke}@speech.sri.com

Abstract

We describe a "direct modeling" approach to using prosody in various speech technology tasks. The approach does not involve any hand-labeling or modeling of prosodic events such as pitch accents or boundary tones. Instead, prosodic features are extracted directly from the speech signal and from the output of an automatic speech recognizer. Machine learning techniques then determine a prosodic model, which is integrated with lexical and other information to predict the target classes of interest. We discuss task-specific modeling and results for a line of research covering four general application areas: (1) structural tagging (finding sentence boundaries, disfluencies), (2) pragmatic and paralinguistic tagging (classifying dialog acts, emotion, and "hot spots"), (3) speaker recognition, and (4) word recognition itself. To provide an idea of performance on real-world data, we focus on spontaneous (rather than read or acted) speech from a variety of contexts—including human-human telephone conversations, game-playing, human-computer dialog, and multi-party meetings.

1. Introduction

Recent years have seen increasing interest in using prosody for speech technology. From a large literature in linguistics and related fields, we know that prosody provides valuable information often not available from text alone—for example information on phrasing and disfluencies, pragmatics and emotion, and even (as is clear from speech heard through a wall) information on speaker identity. It is well understood that humans make ample use of such information in everyday communication. Thus, capturing prosodic knowledge in speech technology is one way to make systems more intelligent and human-like.

To use prosody effectively in automatic systems, however, one needs to address some key challenges. First, one must perform all of the processing automatically—from extracting and normalizing features, to using the features in some way to aid the application. Second, the very words being spoken must be determined automatically and may contain errors—adding to the overall challenge for both lexical cues and for prosodic cues that depend on word locations. Finally, to be viable for many desired applications, the modeling must work not only for read or acted speech (which have formed the basis for much of the descriptive work on canonical prosody) but in particular for spontaneous speech, which is much less regular in prosodic behavior.

Along with colleagues at SRI and more recently at ICSI, we have tried to address these concerns by developing a "direct modeling" approach to incorporating prosody in various speech technology tasks [14]. The approach involves no hand-labeling or modeling of prosodic events such as pitch accents or boundary tones. Instead, prosodic features are extracted directly from the speech signal and machine learning techniques determine the best way to use these features in predicting the target classes of interest.

Although this area of research is currently relatively small, it is expanding, with similar work developed independently (most notably the work of researchers at the University of Erlangen), and other work building on the work described here. This overview will focus mainly on selected studies from the SRI/ICSI line of research, due to space limitations and also because the work is generally representative. We note that many references to interesting work by other researchers in the various application areas covered here can be found within the citations listed in Section 3.

The advantages of direct modeling are both practical and technical. Direct modeling is less costly than modeling via intermediate representations, since no human annotation of prosody is required. Furthermore, direct modeling can lead to better performance, because features are optimized for performance on the end task of interest rather than for the detection of intermediate labels. Direct modeling also avoids the problems of human error or subjectivity in labeling. Finally, direct modeling is more easily ported across tasks, domains, and even languages.

This paper is organized into two main sections. The first section describes the general approach. It includes methods for automatic feature extraction based on the output of a speech recognizer, feature normalization and stylization, machine learning techniques for predicting target classes from prosodic features, and methods for combining prosodic models with lexical and contextual information from statistical language modeling. The second section discusses more detail on work in four selected application areas: (1) structural tagging, (2) pragmatic and paralinguistic tagging, (3) speaker recognition, and (4) word recognition itself. Across studies, data come from a range of corpora of spontaneous speech, including human-human telephone conversations, human-human game-playing, human-computer dialog, and multi-party meetings.

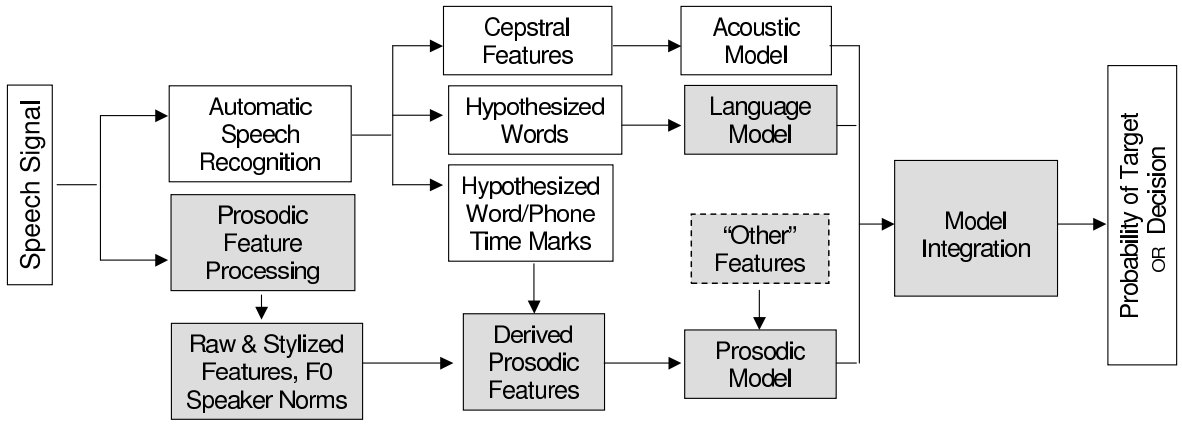

Figure 1: *Schematic diagram of general approach across tasks. White boxes indicate processing in standard speech and speaker recognition. Shaded boxes indicate processing added in the prosody modeling approach.*

2. General method

The overall approach can be summarized as shown in Figure 1. We refer to this figure to provide a high-level description of the common framework across tasks. Further details (in particular on the language model, prosodic model, and integration) are provided by individual application area in Section 3.

2.1. Standard speech and speaker recognition

The white boxes in Figure 1 indicate components that are standard in current automatic speech and speaker recognition systems. *Cepstral features* and the *acoustic model* are included as part of the approach because they are used in the standard speaker recognition systems. For word recognition we use a standard (prosody-unaware) large-vocabulary speech recognizer, which has its own acoustic and language models that are separate from the components shown in the figure. We will use not just the string of word hypotheses but also the corresponding phone and time alignment information. A thorough modeling and processing of prosodic information would make use of the set of possible recognition hypotheses, appropriately weighted. However, to simplify the discussion we will consider only the 1-best recognition hypothesis here, except in Section 3.4 in which the subject is prosody for the word recognition task itself.

2.2. Event language model

Shaded boxes in Figure 1 represent additional processing steps involved in our modeling approach. The *language model* shown captures not only word sequences, but rather the joint pattern of words and target events. Details differ depending on the type of task. For example in modeling information such as sentence boundaries, which can be considered to occur between words, the language model creates N-grams of words and interword-boundary events. In the case of dialog acts however, there are actually two aspects to the language model: the sequence of words within a dialog act, and also the sequence of dialog acts themselves. In the case of speaker recognition, language modeling focuses on N-gram sequences that reflect idiosyncratic word usage patterns.

2.3. Event prosodic model

For most of our work we have used either decision trees or Gaussian mixture models to model prosodic information. Both approaches are well established and have been shown to be useful for a variety of machine learning tasks. Decision tree models are "grown" by asking one question at a time of the available features. The feature queried in each question, as well as the threshold value in the question (e.g., is normalized pitch in the sample above .70?) is that which best distinguishes the classes at that node in the tree. Since our different tasks use different features, separate trees are grown for each task. In the testing phase, the decision tree estimates, for each sample X, the posterior probability of each of the classes C, yielding $P(C|X)$.

For simple tasks these estimates can be used directly, but often we will use the estimated probabilities in combination with other probabilistic knowledge to compute a combined estimate from multiple knowledge sources. Although there exist other classifier architectures and associated learning algorithms that also estimate posterior class probabilities, we have tended to use decision trees because they perform well, can be inspected to gain insight into features, and because good public software is available.

The second kind of statistical model we use for prosodic features is a Gaussian mixture model (GMM). GMMs are linear combinations of multivariate Gaussian distributions that model $P(X|C)$. GMMs can be converted into posterior classifiers using Bayes Rule: $P(C|X) = P(C)P(X|C)/P(X)$, where $P(C)$ is the prior distribution of the classes, and $P(X)$ is the marginal distribution of the data. Thus, GMMs are fundamentally different from decision trees in *what* they model (as well as how), since they estimate the probability of the data given a class, i.e., $P(X|C)$. They are limited to data that can be represented as real-valued feature vectors, and (without modification) cannot deal with datapoints that have missing features. However, they have other advantages, such as the ability to train a model on a large set of data, and then adapt it to new data.

2.4. Prosodic features

The prosodic classifiers take as input a set of human-designed **derived prosodic features** that we hope will be useful for the particular task at hand. **Hypothesized time marks** from the speech recognizer are used to compute and normalize a variety of duration and speaking-rate features. From phone-level timing information, we compute duration features based on phones, vowels only, syllables, and sub-syllable units such as rhymes. To produce meaningful derived duration features, we normalize durations in various ways, using overall statistics for phone, syllable, and state-level information from all data in the training set. For some tasks, we also normalize durations based on local speaking rate and/or duration habits of the particular speaker. For tasks in speaker recognition, we use patterns of unnormalized and word- or phone-specific durations, to a level of detail that includes state (sub-phone) units.

For additional prosodic features we perform **prosodic feature processing** directly on the speech signal. From this we obtain F0, voicing, energy, and spectral tilt information. Also in this step, we apply some useful smoothing and fitting techniques. For example, F0-based features benefit greatly from a post-processing stage that fits linear splines to the raw F0 estimates and models octave errors [16]. We also fit a log-normal tied-mixture model of pitch to a speaker's overall data to obtain estimates of the speaker's F0 baseline, which we have found useful for pitch range normalizations [16, 15].

Derived pitch and energy features are then designed for each task. These include *local estimates of pitch range* (e.g., average or maximum speaker-normalized pitch near a word boundary [for sentence boundaries] or over an utterance [for emotion or hot spots]); *local estimates of pitch contour type and magnitude* (e.g., falling before a word boundary [for sentences] or rising at the end of an utterance [for dialog acts]), and *overall pitch variation patterns* such as contour type and variation over a conversation [for speaker recognition]. Similar features are also computed for energy patterns.

Finally, for some tasks we also make use of **other features** or knowledge sources that interact in interesting ways with our prosodic features. For example, in work on detecting frustration in a corpus of speakers talking with a computer system, it was clear that utterances involving a repeated request (i.e. after a previous machine misunderstanding) were more likely to be frustrated than those involving a new request or response. Thus, a feature capturing "repeat attempts" was also modeled in the prosodic classifier.

2.5. Model integration

As indicated in Figure 1, the prosodic model is combined with various other knowledge sources. We minimally combine prosody with information from the language model. For speaker recognition, we include frame-based acoustic features; for emotion classification, as just noted above, we include information from the discourse context. More details on integration as it pertains to specific tasks is provided in Section 3.

2.6. Coping with classifier greediness and skewed data

Decision trees have two main problems, which we have tried to address. First, to help overcome the problem of greediness, we wrap a feature subset selection algorithm around the standard tree growing algorithm, thereby often finding better classifiers by eliminating detrimental features up front from consideration by the tree [15]. Second, to make the trees sensitive to prosodic features in the case of highly skewed class sizes, we train on a resampled version of the target distribution in which all classes have equal prior probabilities. This approach has additional benefits. It allows prosodic classifiers to be compared (both qualitatively and quantitatively) across different corpora and tasks. In addition, classifiers based on uniform prior distributions are well suited for integration with language models, as described further in the next section.

3. Applications

Having provided a brief overview of the key ideas in our approach to computational prosody, we now summarize some sample applications of the framework.

3.1. Structural tagging

Automatic speech recognizers typically output only a stream of words, which lacks the punctuation, capitalization, and formatting that help to convey structure in written text. In structural tagging, the task is to annotate this simple word stream with basic information related to phrasing, including information on the boundaries of sentence units and larger units such as paragraphs or topics, as well as finding local interruptions of structure including disfluencies. In written language, such phenomena are conveyed via punctuation or formatting. In structural tasks, the events of interest occur at candidate locations that correspond either to specific words or word sequences (as in filled pauses or discourse markers used as fillers), or to inter-word boundaries (as in sentence or topic boundaries, and also as in the interruption point of disfluencies). Prosodic cues are well understood to mark both major phrase boundaries and disfluencies in English, as well as in other languages. Although it is often difficult to find clear prosodic marking of structure in spontaneous speech, stochastic models of prosody, especially when combined with lexical information, do show promise for these tasks.

3.1.1. Modeling

The tasks of sentence segmentation, topic segmentation and disfluency detection can all be cast as word-boundary classification tasks. That is, each location between words is to be classified as a sentence end, topic change, or disfluent interruption point, respectively, versus an unmarked word boundary. For prosodic modeling purposes, each word boundary is represented by a vector of features describing the location of interest. The prosodic classifier then estimates the probability $P(C|F)$ of the boundary type C given that feature vector F. For reasons given earlier we use decision tree classifiers for these tasks.

As one would expect, it is important to combine prosodic cues with information from other knowledge sources. The lexical context, in particular, can convey useful information about the type of word boundary. For example, it is much more likely that a sentence boundary precedes the word "the" than follows it. We model the interaction between words and boundary classes (sentence boundaries, topic boundaries, and interruption points) using N-gram language models. These are statistical models that estimate the probability of a sequence of tokens $P(t_1 t_2 \ldots t_n)$ based on the cooccurrence statistics of up to N consecutive tokens. In our case, the token stream consists of both words and markers for the boundaries of interest. Thus the language model models the cooccurrence of words and boundary types.

Prosodic and word-level knowledge sources can be combined by manipulating probabilities from the language model

and the decision tree. We use C to denote the classes of all boundaries in a stretch of speech, W to denote the corresponding word sequence, and F to denote the corresponding prosodic features. Our goal is to obtain a model for the combination of all three elements $P(W, F, C)$, and then to choose the sequence of boundary classifications that have the highest probability given the observed words and prosody:

$$\underset{C}{\operatorname{argmax}}\, P(C|W, F) = \underset{C}{\operatorname{argmax}}\, P(W, F, C) \quad .$$

The combined model for $P(W, F, C)$ is constructed as follows:

$$\begin{aligned} P(W, F, C) &= P(W, C)P(F|W, C) \\ &\approx P(W, C)P(F|C) \\ &= P(W, C)P(C|F)P(F)/P(C) \end{aligned}$$

The second line is an approximation; it makes the simplifying assumption that the prosodic features of a boundary depend only on the boundary class, not on the words. Of course this assumption is not always true. The phonetic makeup of specific words can be correlated directly with prosodic differences that are difficult to normalize for. Furthermore, there are certain indirect relationships between words and prosody; for example, utterance boundaries are frequently found after backchannels like "yeah" and "uh-huh", but the prosodic nature of these utterance boundaries differs from boundaries of utterances containing semantic content. Despite these cases, however, it is often reasonable to make this independence assumption. In the final step we use Bayes Rule to make use of the decision tree model, which gives us $P(C|F)$. If the decision tree was trained on a resampled training set with equated priors, as suggested in Section 2.6, then the tree posteriors can be used directly in lieu of $P(F|C)$, since the quantities differ only by a constant factor that is independent of C.

A last consideration in this approach is how to find the set of boundaries that maximizes the joint probability $P(W, F, C)$. Here we will again gloss over the details, and just mention that the form of the model described above (especially the fact that it involves an N-gram language model for $P(W, C)$) makes it possible to use efficient search algorithms derived from hidden Markov models [12].

3.1.2. Selected results

We have examined the contribution of prosody to sentence boundary detection in a number of different studies. We have compared a language-model-only approach to an approach that combines language model and prosodic information, and found consistently that the latter performs significantly better. For example, for spontaneous telephone conversations, using true words for both the language and prosody model, prosody provided a significant 7% improvement over the language model alone [15]. The improvement from prosody on this task is even larger if one looks at read speech. In a study of anchor speech in news broadcasts, we found that the prosodic model alone outperformed the language model, despite the fact that the language model was trained on more than an order of magnitude more data than the prosody model. By combining models, we obtained an additional 19% relative error reduction for true words and 8.5% for recognized words over the prosody model alone [15]. We have also found improvements for sentence boundary detection from prosody for data from natural meetings. In this case, despite the speech being spontaneous, prosody alone significantly outperformed the language model alone for the case of recognized words. And again, a further win was obtained by combining the prosodic and lexical models [2]. These results have generally shown a large reliance (as expected) on pause durations, but also the use of durational lengthening and pitch information.

The studies just mentioned have assumed an offline task, which means that prosodic features can be measured both before and after potential boundary locations. In related work, we have also asked whether prosody could be used in an *online* application—to find the ends of utterances to a spoken dialog system, in real time. Currently such "endpointing" is done by waiting for a silence of some fixed threshold duration. This standard approach causes needless waiting time for users at real utterance ends. It also results in premature cutoffs (and angry users!) when speakers pause due to hesitation and expect the system to know (as a human does) that the utterance is not yet finished. We found that the prosody of speech *before* pauses is quite helpful in distinguishing final from hesitation pauses. For example, if the rate of premature cutoffs is held constant, prosody modeling allows for an up to 81% reduction in the average user waiting time [7].

In addition to sentence boundaries, we have found prosody to be helpful in detecting larger-level structures such as topic boundaries in news speech [15]. We have also found it to aid in finding points at which speakers become disfluent [10]. In the case of disfluencies, prosody could be particularly useful for finding disfluency types like false starts, which (unlike filled pauses and repetitions) are difficult to detect using lexical information alone.

3.2. Pragmatic and paralinguistic tagging

In English, and in human languages in general, it is well known that discourse-level, pragmatic and paralinguistic information can be conveyed through prosody. For example, for English it is often said that certain types of questions, such as yes-no questions, are marked with rising intonation. In studies of emotion, variation in a speaker's arousal level and along a positive-negative dimension is associated with prosodic features such as pitch. In spontaneous human-human speech, both dialog act classification and emotion classification are not as straightforward as in restricted or acted domains. For example, in natural conversation, many questions do not have rising intonation, and conversely rising intonation is often used on nonquestions (the notorious "valley girl" intonation that has insidiously crept into a much broader usage).

3.2.1. Modeling

In dialog act tagging, the goal is to classify each utterance as one of a number of dialog act types. The unit of classification is thus a whole utterance. Accordingly, prosodic features are extracted at the utterance level, and a decision tree is trained to compute $P(C|F)$, where C is the dialog act label and F are the prosodic features of the utterance.

Again we have to consider how to combine the prosodic model with other non-prosodic knowledge sources. In dialog act tagging there are two additional types of knowledge that can be modeled. First, the words in an utterance can provide valuable information. We can capture lexical cues with N-gram language models that are specific to each dialog act type. For each dialog act class C, a model is trained to estimate $P(W|C)$, where W are the words in the utterance. By assuming that the words across different utterances are independent once the dialog act types are given (not a valid assumption, but workable

in practice) we can use these language models to model the whole conversation (i.e., let W now denote the word sequence and C the sequence of dialog act labels for the entire conversation). The second additional knowledge source comes from constraints on how dialog acts typically follow each other; e.g., a question is more often followed by a statement than by another question. Such constraints can also be modeled with language models, except that the tokens in question are now dialog act labels instead of words. This is sometimes called a *dialog grammar*, and can conveniently be modeled using N-grams as well. The dialog grammar estimates $P(C)$, the prior probability of a dialog act *sequence*.

Again we search for the class (dialog act) assignment that maximizes the posterior probability given the observed words and prosody:

$$\underset{C}{\operatorname{argmax}}\, P(C|F,W) = \underset{C}{\operatorname{argmax}}\, P(W,F,C)$$

The joint probability of words, prosody, and dialog acts is obtained by decomposition into the models mentioned earlier:

$$\begin{aligned} P(W,F,C) &= P(C)P(W|C)P(F|W,C) \\ &\approx P(C)P(W|C)P(F|C) \end{aligned}$$

In this case we make the simplifying assumption that the prosodic features are independent of the words once the dialog act class is given. As before, the probability of $P(F|C)$ can be derived from the decision tree estimate $P(C|F)$ by applying Bayes Rule. Emotion tagging can be thought of as a special case of dialog act tagging, although we typically simplify the framework in this case by not modeling the sequence of emotions themselves.

3.2.2. Selected results

In the case of dialog act classification, although we have found prosody to be of some help for overall classification [17], it tends to be most helpful for specific class distinctions. One such example, as might be expected, is the difference between statements and questions. In a study of telephone conversations we found that prosodic features, particularly F0, performed better than recognized words alone, and combining the two knowledge sources yielded at further 16% relative reduction in error [13]. We found a similar benefit in distinguishing questions from statements in a corpus in which two speakers collaborated on a simulated military exercise via a video display. In this case, we also investigated the use of unsupervised methods for training models, since we had only limited hand-labeled data. We found that the combined use of prosodic information and unsupervised labeling reduced our tagging error rate by up to 16%, compared to baseline systems using only word information and labeled data [19].

Other distinctions for which prosody is useful involve short responses like "yeah", "right", and "uh-huh", which can serve a variety of different pragmatic functions. For example in the corpus of telephone speech, prosody aided the discrimination of explicit agreements from simple backchannel responses by (yes, yet again) 16% over words alone [13]. In data from multiparty meetings, we looked at differences between four dialog acts types: backchannels, agreements, acknowledgments, and floor-grabbers—all of which can be performed using similar words. Interestingly, by looking at two-way comparisons, we found that quite different prosodic features cue the different distinctions [3].

A second type of utterance-level task for which we have found prosody to be useful is emotion classification. We focused on user frustration in a study of telephone speech to an automatic air travel planning system. Results showed that words alone were poor predictors of emotion in this domain. This is probably because users had to stay within the task vocabulary, and many utterances like "no" could be used in a variety of contexts. Compared with a lexical-only model, classification error was reduced by roughly 14% relative for a prosody model alone, and by roughly 27% relative using prosody model that included a "repeated-attempt" feature in the decision tree building [1]. In related work, we have begun to look at locations we refer to as "hot spots", or points of high participant "involvement" in data from multi-party meetings. We have found that automatically extracted pitch and energy features show significant correlations with involvement level—a finding that could aid applications such as automatic browsing or summarization [21].

3.3. Speaker Recognition

A common task in speaker recognition is verification: determining whether a speech sample comes from a known target speaker, or from someone else (possibly an intentional imposter). Many commercial systems use verification for user access purposes. Other applications include methods for searching large archives of conversational audio data to find voice data from speakers of interest to intelligence or law enforcement agencies.

Conventional speaker recognition systems use distributions of spectral cues from very short and essentially unordered time slices of speech. For applications in which more than a few seconds of speech are available to train and to test a system—such as the tasks involving flagging of large databases of conversations—we and others have recently proposed that longer-term features such as prosodic cues could provide additional speaker-specific information.

3.3.1. Modeling

The standard decision paradigm for speaker verification is to evaluate the observed speech features F against two models: one model for the target speaker $P(F|\text{target})$, and another model for the likelihood that the features could have been generated from a generic speaker: $P(F|\text{generic})$. The latter model is called a *background model*. The two models likelihoods are then compared. If their ratio $P(F|\text{target})/P(F|\text{generic})$ exceeds a threshold, the system accepts the speech sample as coming from the target speaker. The choice of threshold is flexible: by raising it the system can lower the probability of false acceptance, at the expense of frequent rejections of the true speakers. Conversely, a low threshold minimizes the chance of false rejections, but also incurs more false acceptances. The quality of a system is therefore often summarized by the *equal error rate* (EER), the point at which false acceptances and rejections occur with equal frequency.

For various reasons the target speaker models are not trained from scratch, but rather by *adapting* the background models. This involves starting with the background model and then adjusting its parameters to better represent the target speaker, but without "forgetting" the statistics of the background speaker population (which typically involves hundreds or thousands of speakers). One important result of this procedure is that the likelihoods of the background and target models are numerically comparable over a wide range of possible ob-

served speech samples, which allows for a meaningful comparison against a single threshold.

The need to estimate likelihoods and adapt models makes GMMs a convenient tool. We extract real-valued feature vectors for each speaker. These features can be extracted at the level of the speaker, utterances, words, or whatever unit of speech is convenient or natural for a given feature. In the case of multiple samples per speaker (such as when the features are word-based) the sample likelihoods are simply multiplied to form an overall score. This implies an assumption that the observations are independent, which is not true, but which seems to be a viable approximation in practice.

Features are extracted for the background training corpus, and a background GMM is estimated. For each target speaker, the background model is adapted to form a target speaker model. On a given test speaker the ratio of the two model likelihoods is formed. To find the EER, all possible threshold values are checked to locate the setting that equates false acceptances and false rejections. As with other tasks, one usually wants to combine prosodic knowledge with other, more standard knowledge sources. In this case the likelihood ratios for different speaker features are combined using several standard techniques. For example a linear combination of the log likelihood ratios (with empirically optimized weights) is commonly used, or a non-linear combination via a neural network.

3.3.2. Selected results

We have studied the use of long-range prosodic features through participation in the NIST 2003 Speaker Recognition Evaluation extended data task [11]. We evaluated the contribution of prosody by comparing it to our best nonprosodic system, which consists of a state-of-the-art frame-based cepstral system combined with a system based on lexical N-grams (after [5]). The EER for this combined system is 0.57%.

One prosodic feature type we have found to be extremely useful is duration, particularly when duration is constrained by segmental information. We create three different systems: a word-based system (sequence of phone durations in specific words), a single-component phone-based system (durations of phones themselves, regardless of location), and a 3-component phone-based system (sequence of durations of the states within the phones). These vectors are then modeled by GMMs. By combining these duration systems with each other and then with the baseline+lexical system above, we reduce the EER from 0.57% to 0.29%—a nearly 50% relative reduction in error [6, 9].

We have also begun exploring a range of other prosodic features, which we refer to as "NERFs" (for New Extraction Region Features). These features are delimited not by conventional units such as frames, words, or phones, but rather by the presence of various automatically detectable events that we hope may correspond to meaningful units in some way. Within each new extraction region, a vector of different features is defined and then modeled using GMMs. For example, we have found that features such as maximum stylized pitch and maximum phone-normalized duration are useful when extracted from a region delimited by pauses larger than 500 milliseconds. Using a system based on features in this "pause-to-pause" region, we can reduce the baseline EER by about 15%, an encouraging preliminary result especially given the low density of NERFs compared with features modeled by other systems [9].

3.4. Word recognition

The applications discussed so far have all aimed to add some type of tagged information (e.g., sentence boundaries, dialog acts, emotion, speaker information) to the output of an automatic speech recognizer. We conclude this paper by taking a look at word recognition itself. There are two ways in which prosody could be used to improve word recognition. The first way is develop models that capture prosodic information about words themselves. There has been some successful work in this area for spontaneous speech, involving duration patterns, which is discussed below. The second way to potentially improve word recognition through prosody is to apply prosody to tasks such as the structural and other tasks described earlier, and then use the matches between acoustics and tags, and tags and words, to help determine the most likely word sequence.

3.4.1. Modeling

The goal in word recognition is to find the word string W that has the highest posterior probability given a set of acoustic observations X. Using Bayes Rule, this is usually expressed as maximizing the product of a prior word string probability $P(W)$ (the recognizer language model) and an acoustic likelihood $P(X|W)$ (the recognizer acoustic model).

Some researchers have built a more detailed version of both acoustic and language models by replacing the word labels with a vocabulary that distinguishes phonological prosodic events, such as stress and boundary tones. For example, in the work by Chen et al. [4], each word comes in different versions depending on whether or not it is stressed, and whether or not it precedes a prosodic phrase boundary. Let us denote these prosodic labels with L. The acoustic model can now be conditioned on the prosodic events, $P(X|W, L)$, allowing it to capture how stress and phrasing affect the spectral properties of the speech signal. The new language model $P(W, L)$ is also potentially more accurate, since words with different prosodic characteristic might well have different cooccurrence statistics with surrounding words. A fringe benefit of this approach is that the recognizer outputs not just words, but also prosodic tags. I.e., it finds

$$\operatorname*{argmax}_{W,L} P(W, L|X) = \operatorname*{argmax}_{W,L} P(W, L)P(X|W, L)$$

The biggest drawback to such an approach, as noted earlier for approaches involving intermediate categories, is that it requires training data that is labeled for the prosodic distinctions used by the models. Given that state-of-the-art recognizers use hundreds or even thousands of hours of training data, and that prosodic labeling is notoriously difficult, this is a significant limitation. It is however possible that automatic prosodic labeling will become accurate enough at some point to support this approach on a larger scale.

Our work on leveraging prosody in word recognition involves using features and events that can be directly extracted from the recognizer output or that are a by-product of recognition. This involves defining a prosodic observation stream F in addition to the standard spectral features X. The goal of the recognizer then becomes

$$\begin{aligned}\operatorname*{argmax}_{W} & P(W|X, F) \\ &= \operatorname*{argmax}_{W} P(W)P(X, F|W) \\ &= \operatorname*{argmax}_{W} P(W)P(F|W)P(X|W, F)\end{aligned}$$

The new model component $P(F|W)$ captures the dependence of prosodic observations on the hypothesized words. Gadde [8] describes one instantiation of this approach, in which phone durations (normalized for rate-of-speech) in each word serve as the features F, and $P(F|W)$ is estimated by GMMs. Vergyri et al. [20] extend this approach by also modeling the pauses (and their lengths) between words.

Of course the first approach (adding prosodic labels to the vocabulary) can be combined with the second (adding prosodic features as observations). In fact, [4] employ their stress and phrase-conditioned models together with a frame-level F0 feature in the acoustic model $P(X, F|W, L)$, where F is the F0 feature.

Yet another variant employs a standard vocabulary and acoustic model (thereby avoiding the need to label large amounts of training data prosodically), but models the effect that linguistic structures beyond the words have on the directly extracted prosodic features F. We let S denote the linguistic structure: for example, the location of sentence boundaries, or syntactic parses. We can model the effect that this structure has on prosody, via a model $P(F|S, W)$. We also need to characterize how the structures "go along" with different words strings, i.e., a model for $P(S|W)$. Because these model components are separate from the standard acoustic and language models they can be trained on smaller amounts of data, or on data that has been automatically annotated for the structures S. Once all these components are in place they can be used as follows:

$$\begin{aligned}
&\operatorname*{argmax}_W P(W|X, F) \\
&\quad = \operatorname*{argmax}_W P(W)P(X, F|W) \\
&\quad = \operatorname*{argmax}_W P(W)P(X|W, F)P(F|W) \\
&\quad = \operatorname*{argmax}_W P(W)P(X|W) \\
&\qquad\qquad \sum_S P(F|S, W)P(S|W)
\end{aligned}$$

The last line implies that we do not try to extract the single best structure hypothesis for a given word hypothesis, but instead sum over all possible structures. This leads to more accurate overall results and helps to mitigate possible deficiencies in the models $P(F|S, W)$ and $P(S|W)$. We have explored this last approach successfully by considering sentence boundaries and disfluent interruption points as the structural elements underlying the words [18]. This allowed us to reuse prosodic models previously used for sentence segmentation and disfluency detection for the model component $P(F|S, W)$. Similarly, the N-gram language models used in those tasks can be employed to estimate $P(S|W)$.

3.4.2. Selected results

As reported in [20], the explicit phone duration modeling with GMMs and pause duration prediction with language models reduced the word error rate on a conversational recognition speech task by 2.1% relative. The modeling of hidden sentence boundaries and disfluencies provided a 1.7% relative error reduction. When both techniques are combined, they reduce error by 3.1%. This shows that, as expected, the two models are not completely orthogonal, since they both model related pause and duration aspects. Although the improvement on spontaneous speech is relatively small, it is still highly significant.

4. Conclusions

We have briefly outlined an overall approach for direct modeling of prosody for various speech technology applications. The approach extracts features from the speech signal and from the associated output of an automatic speech recognizer, and models those features using either decision trees or Gaussian mixtures. Information from the prosodic model is further combined with information from lexical features (and, in some cases, additional knowledge sources) to predict the target classes of interest.

The approach is completely automatic and has proven successful when applied to spontaneous speech data for a range of applications, including structural tagging, pragmatic and paralinguistic tagging, speaker tagging, and word recognition. While the overall framework is common across tasks, details of the features, modeling, and integration are task-dependent. In particular, different tasks require different prosodic feature types and different regions from which to extract and define the features. Tasks also differ in terms of the relative contribution of prosodic versus lexical features to overall performance.

In conclusion, we would like to emphasize that this brief overview has provided only a small sample of work in this new and interesting area. We hope that in the long term, further work on automatic prosody modeling in the greater research community will help to make speech technology become just a little bit more "human".

5. Acknowledgments

This research was supported by NSF (through a STIMULATE award, an ITR award, and a KDD award); by DARPA (through TRVS, Communicator, ROAR, SPINE Seedling, and EARS awards), and by NASA contract no. NCC 2-1256. The views herein are those of the authors and do not reflect the views or policies of the funding agencies.

6. References

[1] J. Ang et al. Prosody-based automatic detection of annoyance and frustration in human-computer dialog. In J. H. L. Hansen and B. Pellom, editors, *Proc. ICSLP*, vol. 3, pp. 2037–2040, Denver, 2002.

[2] D. Baron et al. Automatic punctuation and disfluency detection in multi-party meetings using prosodic and lexical cues. In J. H. L. Hansen and B. Pellom, editors, *Proc. ICSLP*, vol. 2, pp. 949–952, Denver, 2002.

[3] S. Bhagat et al. Automatically generated prosodic cues to lexically ambiguous dialog acts in multiparty meetings. In *Proceedings International Congress of Phonetic Sciences*, pp. 2961–2964, Barcelona, 2003.

[4] K. Chen and M. Hasegawa-Johnson. Improving the robustness of prosody dependent language models based on prosody syntax dependence. In *Proceedings IEEE Workshop on Speech Recognition and Understanding*, pp. 435–440, St. Thomas, U. S. Virgin Islands, 2003.

[5] G. Doddington. Speaker recognition based on idiolectal differences between speakers. In P. Dalsgaard et al., editors, *Proc. EUROSPEECH*, pp. 2521–2524, Aalborg, Denmark, 2001.

[6] L. Ferrer et al. Modeling duration patterns for speaker recognition. In *Proc. EUROSPEECH*, pp. 2017–2020, Geneva, 2003.

[7] L. Ferrer et al. A prosody-based approach to end-of-utterance detection that does not require speech recognition. In *Proc. ICASSP*, vol. 1, pp. 608–611, Hong Kong, 2003.

[8] V. R. R. Gadde. Modeling word durations. In B. Yuan et al., editors, *Proc. ICSLP*, vol. 1, pp. 601–604, Beijing, 2000. China Military Friendship Publish.

[9] S. Kajarekar et al. Speaker recognition using prosodic and lexical features. In *Proceedings IEEE Workshop on Speech Recognition and Understanding*, pp. 19–24, St. Thomas, U. S. Virgin Islands, 2003.

[10] Y. Liu et al. Automatic disfluency identification in conversational speech using multiple knowledge sources. In *Proc. EUROSPEECH*, pp. 957–960, Geneva, 2003.

[11] National Institutes of Standards and Technology. The NIST year 2003 speaker recognition evaluation plan. http://www.nist.gov/speech/tests/spk/2003/doc/2003-spkrec-evalplan-v2.2.pdf, 2003.

[12] L. R. Rabiner and B. H. Juang. An introduction to hidden Markov models. *IEEE ASSP Magazine*, 3(1):4–16, 1986.

[13] E. Shriberg et al. Can prosody aid the automatic classification of dialog acts in conversational speech? *Language and Speech*, 41(3-4):439–487, 1998.

[14] E. Shriberg and A. Stolcke. Prosody modeling for automatic speech recognition and understanding. In M. Johnson et al., editors, *Mathematical Foundations of Speech and Language Processing*, vol. 138 of *IMA Volumes in Mathematics and its Applications*, pp. 105–114. Springer, New York, 2004.

[15] E. Shriberg et al. Prosody-based automatic segmentation of speech into sentences and topics. *Speech Communication*, 32(1-2):127–154, 2000. Special Issue on Accessing Information in Spoken Audio.

[16] K. Sönmez et al. Modeling dynamic prosodic variation for speaker verification. In R. H. Mannell and J. Robert-Ribes, editors, *Proc. ICSLP*, vol. 7, pp. 3189–3192, Sydney, 1998. Australian Speech Science and Technology Association.

[17] A. Stolcke et al. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–373, 2000.

[18] A. Stolcke et al. Modeling the prosody of hidden events for improved word recognition. In *Proc. EUROSPEECH*, vol. 1, pp. 307–310, Budapest, 1999.

[19] A. Venkataraman et al. Training a prosody-based dialog act tagger from unlabeled data. In *Proc. ICASSP*, vol. 1, pp. 272–275, Hong Kong, 2003.

[20] D. Vergyri et al. Prosodic knowledge sources for automatic speech recognition. In *Proc. ICASSP*, vol. 1, pp. 208–211, Hong Kong, 2003.

[21] B. Wrede and E. Shriberg. Spotting “hotspots” in meetings: Human judgments and prosodic cues. In *Proc. EUROSPEECH*, pp. 2805–2808, Geneva, 2003.