



# Modeling Phrasing and Prominence Using Deep Recurrent Learning

Andrew Rosenberg<sup>1</sup>, Raul Fernandez<sup>2</sup>, Bhuvana Ramabhadran<sup>2</sup>

<sup>1</sup>Queens College (CUNY)

<sup>2</sup>IBM TJ Watson Research Center, Yorktown Heights, NY – USA

andrew@cs.qc.cuny.edu, {fernandra, bhuvana}@us.ibm.com

## Abstract

Models for the prediction of prosodic events, such as pitch accents and phrasal boundaries, often rely on machine learning models that combine a set of input features aggregated over a finite, and usually short, number of observations to model context. Dynamic models go a step further by explicitly incorporating a model of state sequence, but even then, many practical implementations are limited to a low-order finite-state machine. This Markovian assumption, however, does not properly address the interaction between short- and long-term contextual factors that is known to affect the realization and placement of these prosodic events. Bidirectional Recurrent Neural Networks (BiRNNs) are a class of models that overcome this limitation by predicting the outputs as a function of a state variable that accumulates information over the entire input sequence, and by stacking several layers to form a deep architecture able to extract more structure from the input features. These models have already demonstrated state-of-the-art performance on some prosodic regression tasks. In this work we examine a new application of BiRNNs to the task of classifying categorical prosodic events, and demonstrate that they outperform baseline systems.

**Index Terms:** prosodic analysis, recurrent neural networks,

## 1. Introduction

Speech prosody is a suprasegmental phenomenon that impacts the acoustics of speech over a much longer time range than phones or frame-level acoustics. The presence of a pitch accent (prominence) is defined by the perception of a word or syllable as standing out from its surrounding context, and the perception of phrasing is based on the degree of perceived disjuncture between words. Moreover, some prosodic variation, e.g. that relating to discourse context, can span multiple phrases.

As temporal context is inherent to the communication of information via prosody and the perception of prosodic variation, any successful modeling of prosodic events needs to include some representation of context. This is usually accomplished in two ways: by having the inputs to the model explicitly include some notion of context, or by employing a model that attempts to model contextual variation directly. The former is addressed by extracting acoustic features from regions that span multiple words, or syllables [1]. There is certainly room for improvement in how acoustic features successfully encode appropriate context for modeling; insights in this direction will continue to come from speech science and hearing research. Sequential models, such as HMMs [2] and CRFs [3], on the other hand, incorporate context into the model itself. Both have been previously applied to prosody modeling tasks, with the latter

yielding very strong results. There are two major limitations, however, of the way these have been used to recognize prosodic events. First, they incorporate only previous context. This limits the ability to forward context to inform the decision process. Phrase boundaries are determined by the relationship the acoustic context surrounding a candidate boundary, and pitch accents are determined by their acoustics relative to the previous and following words and syllables. Second, the amount of context used by the model must be specified by the user *a priori*. While it is well understood that surrounding context is critical for modeling prosodic information, the optimal amount of context and the best way to incorporate that into sequential models do not share similar consensus.

A class of models that addresses these limitations is the Bidirectional Recurrent Neural Network (BiRNN), a type of neural architecture containing recurrent hidden layers that can capture, relevant structure from the input at arbitrary time lags. These models naturally incorporate past and future contextual dependencies, and are most successful when the hidden layers consist of Long Short-Term Memory units, cells that can store information from the input until such time as they become relevant to learning. We discuss the strengths of the BiRNN in more detail, especially drawing contrast to the CRF, in Section 5. Motivated by this observation, we apply BiRNNs to modeling prominence and phrasing, as defined under the ToBI model of American English prosody [4] as Pitch Accents and Intonational Phrase boundaries. In both cases, we evaluate the capacity of the BiRNN to model pitch accents and boundary tones as binary detection tasks, comparing performance to the CRF.

The rest of the paper is structured as follows. In Section 2 we describe related work on this task. Section 3 presents the data we use in this work. Section 4 describes and motivates the lexical and acoustic features explored. We present the modeling approaches in Section 5. Section 6 describes the experiments and presents results. Finally we conclude in Section 7.

## 2. Related Work

There is a wealth of research on modeling prosody, in general, and detecting prominence and phrase boundaries, specifically, much of which incorporates context at the feature level. Levov investigated context for recognizing prosodic events and in distinguishing tones in Mandarin [1]. Rosenberg looked at normalizing features by their surrounding context in the detection and classification of prosodic events [5, 6, 7]. Mishra et al. [8] developed a number of new features to represent the shape of an acoustic contour with respect to its context. However, all of this work uses static models to recognize prosodic events.

A number of approaches have been proposed to use sequential models to analyze prosody. Ananthakrishnan et al. [2] used

a Coupled HMM model to predict prominence and phrasing at the syllable and word levels. In general frame-level sequential models do not perform as well as word- or syllable-based models. For example, Wightman decoded word-level hypotheses using a tone-sequence model [9]. The CRF has been used by a number of authors. Gregory and Altun [10] used a CRF with linguistic features to assign prosody on Switchboard data. In previous work, the authors of this paper used CRF modeling to perform phrase assignment from text in a variety of domains [11]. Fernandez and Ramabhadran [3] recognized pitch accents evaluating the impact of the amount of training data on performance. Directly related to this work, Fernandez et al. [12], recently used a BiRNN to assign specific, frame-level prosodic parameters for speech synthesis from lexical features.

### 3. Material

The corpus we are using for this work is a set of approximately 3,730 utterances recorded by a single professional female speaker of US English, which was been annotated using the full ToBI framework by an experienced ToBI annotator. No independent set of annotations, and therefore inter-rater agreement, is known for this corpus. The total material consists of 60,061 word tokens. Since we are using sequential models, we need to preserve the full sequential structure of each utterance, and therefore partition the corpus into disjoint training, development and test set utterance-wise, using an 80%/10%/10% split. This results in token counts of 47,893 for training, 6,268 for development, and 5,900 for the test sets, respectively. Development data is used for early-stopping when training BiRNN models and for all parameter tuning.

### 4. Features

In this section, we describe the word-level features used as predictors of prosodic events, many of which have been used in previous prosody modeling work [5, 3, 11].

#### 4.1. Lexical Features

We extract a relatively small set of lexical features, including the degree of coupling between two words using forward and reverse bigram language models (LMs) to calculate the following two features:  $lm1_k = p(w_k|w_{k-1})$  and  $lm2_k = p(w_k|w_{k+1})$ . These LMs are trained using 8 different corpora from a variety of styles (broadcast news, transcribed conversational speech, etc.) with a vocabulary of approximately 86,000 words. Models are trained on each sub-corpus independently, using Kneser-Ney smoothing, with the final LM is constructed by interpolating the 8 models with weights chosen to minimize perplexity on the LM training data.

Parts-of-speech (POS) are extracted using the Stanford Parser [13]. For lexical items that are split by the parser into several tokens with their respective POS (e.g., contractions such as “don’t”), we maintain only the first tag in order not to disturb the sentence word count, and their respective alignment with other word-level features and labels. We also extract the identity of any following punctuation. Finally, we include 5 Boolean features indicating whether each word is a member of the following broad word-classes: 1) auxiliary verbs, 2) conjunctions, 3) function words, 4) wh-words, 5) adpositions.

#### 4.2. Acoustic Features

All acoustic features used in these experiments are extracted using AuToBI [6]. We use the word—rather than syllable or other region—as the unit of analysis for these experiments, and aggregate any lower-level features (e.g., frame based) for each word

in the data set using available phonetic- and lexical alignments. (cf. Section 3). This choice of unit is consistent with the ToBI model of intonation, which associates prosodic events with the word on which they are realized. Moreover, previous work has shown that it is more reliable to recognize prominence on word regions, rather than syllable or vowel regions [14].

The majority of acoustic features are constructed by first calculating a short (10ms) frame-based acoustic contour, followed by subsequent aggregation of the contour over each word. The extracted base contours are informed by known acoustic correlates of prosodic variation.

**Pitch** We calculate the pitch contour using an implementation of the RAPT algorithm [15] to extract F0. We represent the pitch contour in units of log Hz for two reasons: it is more consistent with human perception of pitch, and Hz values in speech are log-normally distributed. Not all frames contain pitch information due to unvoiced phones, we linearly interpolate the log F0 contour over non-silent frames that have no hypothesized log F0 data. (Silence is determined by an intensity threshold of 30dB.) During silent frames, the pitch value is undefined.

**Intensity** The intensity contour is calculated as the energy in each frame surrounded by a 40ms Hamming window in dB units based on a reference value of  $2 \cdot 10^{-5}$ . In addition to the total energy in the signal, we also extract the energy in a range of non-overlapping frequency sub-bands on the Bark scale, starting at 0 and ending at 22 Bark (the Nyquist rate of the signal files), with a 2-Bark width.

**Pitch/Energy Interaction** Prosody is often a function of the interaction and timing of both pitch and intensity. To compactly represent this interaction, we construct a frame-based contour containing the product of log F0 and intensity. During silent frames, this product value is undefined.

**Spectral Tilt** Spectral tilt, or spectral emphasis captures increased high-frequency energy. This has been associated with increased sub-glottal pressure, and perceptions of prominence [16]. We calculate the spectral tilt as the ratio of energy above 500Hz to the overall energy in the frame.

**Fundamental Frequency Variation** In part to address the problem of missing pitch frames, Laskowski, et al. developed the fundamental frequency variation (FFV) spectrum [17]. This calculates the instantaneous change in fundamental frequency by inspecting the harmonic dilation or contraction in short windows preceding and following a frame. This measure is defined at all points, whether or not the frames contained periodic information. Rather than taking the maximum value of the FFV spectrum (corresponding to the best estimate of instantaneous change), following the prior work on this, we calculate a 7-element FFV spectrum describing the strength of hypotheses that the fundamental frequency is falling (at 3 different rates), staying (roughly) stable, or rising (at 3 different rates).

**Deltas** We also calculate the delta of each of the above contours to capture their change over time.

Based on these contours we calculate a number of aggregations to represent the prosodic content of each word. These include the minimum, maximum, mean ( $\mu = \frac{1}{n} \sum x$ ), standard deviation ( $\sigma$ ) and the z-score ( $(x - \mu) / \sigma$ ) of the maximum within the context of the current word. In addition, we calculate the area under the curve (AUC) ( $\sum x$ ). We also use a number of techniques to represent the shape of the contour within the word. These include Tilt coefficients [18], the Tonal Center of Gravity [19], and the relative likelihoods that the curve is

rising/falling/peak/valley based on isotonic regression [8]. We also reuse the isotonic regression shape modeling to identify the relative position of any discovered peak or valley.

In addition to these aggregations of contour information over the word, we also incorporate the context surrounding the word. For log F0, intensity and spectral tilt, we calculate the z-score-normalized value of the maximum and mean within the word, based on mean ( $\mu$ ) and standard deviation ( $\sigma$ ) statistics calculated from a window that covers the two preceding and two following words. This represents the value of each feature relative to its immediately surrounding words.

There is distinguishing information immediately prior to phrase boundaries, which include pitch and intensity changes corresponding to edge tones, or pre-boundary lengthening. To capture this, we also extract the above features from only the final 200ms of each word.

While pitch accent detection is most reliably performed at the word level, the acoustic excursion associated with prominence is localized on the lexically stressed syllable. To take advantage of this, in addition to extracting these features from each word, we also extract them from the longest syllable in the word. The syllable is identified using an implementation of the pseudo syllabification algorithm described in [20].

The previous acoustic features have all been extracted from each of the acoustic contours. In addition, we extract a few features that are specific to representing the interaction of pitch and intensity [5]. These include 1) the ratio between the areas under the energy and pitch contours 2) the ratio between the relative locations of the pitch and energy peak/valley and 3) the RMSE between the log F0 contour and a one-tenth scaled intensity contour. The units of log Hz and 1/10 dB fall in approximately the same range during normal speech. Lastly, we extract the ratio of voiced to unvoiced frames in the word, the duration in seconds, and the length of preceding and following silences both as a continuous value in seconds, and as a binary indicator feature.

## 5. Modeling approaches

In this section we provide an overview of the two approaches we contrast in this work and use to generate the results reported in Section 6: Bidirectional Recurrent Neural Network (BiRNNs) and Conditional Random Fields (CRFs). CRFs, in particular, have been previously reported to provide state-of-the-art recognition rates when modeling discrete prosodic sequences, and, for this reason, we adopt them as what we think is a fairly strong baseline. After a brief overview, we provide some further comparisons between these two systems which better motivate the use of BiRNNs for the prosody prediction task.

**Bidirectional Recurrent Neural Networks with Long Short-Term Memory Neurons** Recurrent Neural Networks (RNNs) offer a way to take advantage of recent advances to Neural Network based modeling (so-called “Deep Learning”) while incorporating contextual information into the learning process. RNNs are very effective at modeling sequences, generating state-of-the-art performance on sequence modeling tasks including handwriting recognition [7] and phoneme recognition [8]. In the RNN, contextual information is incorporated into the learning process by defining hidden unit activations as a function of both the input at a given time *and* a previous hidden state activations. A Bidirectional RNN (BiRNN) [21] constructs two independent layers of hidden units to accumulate contributions from previous and future “histories” to model both forward and backward dependencies. This type of structure is consistent with the suprasegmental nature of prosodic variation where sur-

rounding context, in both directions, are necessary for appropriate interpretation of prominence and phrasing. The hidden layers are stacked together as in a traditional neural network allowing for deep structure to be learned. The output of a BiRNN is a composition of the forward and backward hidden layers.

Training RNNs (whether uni- or bi-directional) with standard, memory-less units (like tanh or sigmoid) has proven to be particularly difficult, in part due to the vanishing gradient problem, an issue that generally affects learning in deep neural networks when using algorithms such as gradient descent. As gradients estimated through back-propagation accumulate over several layers of connected units, their magnitude tends to shrink. In RNNs, the number of layers over which these must be accumulated is proportional to the number of time units, and, as this can be quite large, the problem is exacerbated. One proposal to alleviate this issue is the use of a compound unit known as a Long Short-Term Memory (LSTM) unit [22, 23].

LSTM units attempt to implement three operations via three internal gates (known as the input, forget, and output gates), which determine 1) whether an input is relevant enough to be stored, 2) how long it should be stored for, and 3) when a stored value should be output. These operations, therefore, provide the unit with some sort of memory state that can be used to withhold information relevant to the learning criterion for an arbitrarily long (or short) amount of time. The learning process adjusts the internal parameters of the LSTM unit to facilitate the three operations described. A more in-depth description of LSTM operation can be found in the references cited.

In our experiments we use three sets of bidirectional hidden layers, with 30, 20 and 10 units respectively. Softmax activation units are used on the output layer, as is standard for classification tasks, and the BiRNN is trained by optimizing cross-entropy using stochastic gradient descent with momentum. After each epoch of training, the cross-entropy and classification error on the development set is evaluated. The training continues until *neither* the development set cross-entropy *nor* classification error decreases for 15 epochs. When this occurs, the model which generated the best development set classification error is selected. For the experiments reported here, we made use of the RNNLib Toolkit [24].

**Conditional Random Fields** A Conditional Random Field (CRF) [25] is an undirected graph with nodes  $x$  and  $y$ , corresponding, respectively, to an observation sequence and a sequence to be inferred, which directly encodes the conditional distribution  $p(y|x)$  using a log-linear model. The most common CRF architecture is a linear-chain CRF. This a structure that assumes that, conditioned on  $x$ , the dependencies of the elements of  $y$  form a Markov chain. In this paper, we make use of this assumption and restrict ourselves to first-order chains with a the conditional distribution of the form:

$$p(y|x) = \frac{\exp\{\sum_{k=1}^K \lambda_k f_k(y_t, y_{t-1}, x_t)\}}{\sum_y \exp\{\sum_{k=1}^K \lambda_k f_k(y_t, y_{t-1}, x_t)\}} \quad (1)$$

where each  $f(y_t, y_{t-1}, x_t)$  is a feature function associated with the token at time  $t$ :

$$f(y_t, y_{t-1}, x_t) = \delta(y_t = i)\delta(y_{t-1} = j)att^l(x_t). \quad (2)$$

and  $att^l(x)$  stands for a Boolean function that signals when some arbitrary property  $l$  of the input sequence is true. For the reported experiments, we have made use of the CRF++ toolkit to estimate the parameters  $\lambda_k$  of the model and decode the test input sequences [26].

Defining features as in 2 is very flexible since this equation allows us to encode arbitrary properties of the input to help predict the output. It can, however, lead to a large number of features with few tokens observed in the training set. To avoid overfitting, any generated feature with fewer than  $f$  occurrences is dropped from consideration at training time. Training the CRFs involves an optimization function which is the sum of the likelihood of the data, plus a regularizer penalizing large weights. A free parameter  $C$  controls the contribution of each of these components to the criterion. We tune values of  $f$  and  $C$  in experiments on the development data, exploring  $f$  at integers  $\in [3, 15]$  and  $C$  at 0.005, at steps of 0.001  $\in [0.001, 0.01]$ , steps of 0.01  $\in [0.01, 0.1]$ , and steps of 0.1  $\in [0.1, 1.5]$ .

**BiRNNs vs CRFs** Both of these models provide sequence-modeling context-dependency capabilities, although they approach the problem in crucially different ways. The linear-chain CRF models label sequences through the explicit use of a first-order Markov relation between labels  $y_t$  and  $y_{t-1}$ , and incorporates arbitrary context dependency by the use of Eq. 2. This is a powerful construct because it allows us to encode arbitrary (local, global, etc.) attributes of the (full) input sequence in the feature definition. Eqn. 2 gives us a mechanism whereby if we know how short- and long-term properties of the input sequence (represented by the raw measures we described in the previous section) interact to contribute to the current prediction, we can hand-craft indicator features to capture that information. The challenge is that for prosody predictions these interactions are not always well understood. This is the limitation that motivates us to suggest the use of RNNs for this task. In RNNs, the full input sequence is presented to the learning model, and through the learning process, the weights of the model are adjusted to try to sift out what is the structure in the input sequence that contributes to the current label prediction. The model is tasked with figuring out how context contributes to the classification. This is a difference that we believe motivates their use for a task such as prosody modeling, where these interactions are not always clear, and the reason for why we choose the comparison between these two particular models in this paper.

**A word about inputs** Both modeling approaches operate on the same raw inputs described. However, RNN models require all features to be numeric, while CRFs operate on indicator features derivable from some discrete property of raw inputs. To address these differences, the data is processed as follows. For RNN training, each of the categorical lexical features is re-encoded with a one-hot vector. For the CRF experiments, we quantize any continuous features into 4 bins, according to this procedure. The continuous LM log-likelihood scores follow a normal distribution. Thus we normalize them based on a Gaussian parameterization. We calculate the mean and standard deviation over the training data, and take the z-score of each observation. From this z-score, we construct four features:  $z \leq -1 \rightarrow \text{VERYLOW}$ ;  $-1 < z \leq 0 \rightarrow \text{LOW}$ ;  $0 < z \leq 1 \rightarrow \text{HIGH}$  and  $z > 1 \rightarrow \text{VERYHIGH}$ . The acoustic features, on the other hand, do not confirm the assumption of normality. We thus use a rank-based normalization, calculating the boundaries of four evenly sized bins ( $q_i[1], q_i[2], q_i[3]$ ) for each variable  $i$  based on the training data. We then assign each observation  $x_i$  of variable  $i$  to one of the four classes  $x_i \leq q_i[1] \rightarrow \text{VERYLOW}$ ;  $q_i[1] < x_i \leq q_i[2] \rightarrow \text{LOW}$ ;  $q_i[2] < x_i \leq q_i[3] \rightarrow \text{HIGH}$  and  $x_i > q_i[3] \rightarrow \text{VERYHIGH}$ .

Occasionally features are undefined, for example, pitch values may be missing or a region may only contain a single point, leading to undefined  $\sigma$  values. When this happens, the undefined value is replaced by the mean value in the training data.

## 6. Experiments

In each of these experiments, we compare the performance of BiRNN models to Conditional Random Fields (CRFs).

Results of Pitch Accent detection experiments are reported in Table 1. In the test data, 1,720 of 5,900 tokens are accented

Model	ACCENTED $F_1$ (P,R)	Accuracy	Error
CRF	0.790 (0.780, 0.799)	87.88%	12.12%
BiRNN	0.811 (0.809, 0.814)	89.03%	10.97%

Table 1: *Pitch Accent Detection results in Accuracy, Error and F-measure for the ACCENTED class.*

yielding an accent rate of 29.15% and a majority class baseline of 70.85%. Due to the skewed class distribution,  $F_1$  is reported to assess performance. On this task, we find a relative reduction of error of **9.48%** when using BiRNN over the CRF models; this corresponds to an absolute  $F_1$  improvement of 2.1% with similar gains to *both* precision and recall.

Results of Intonational Phrase Boundary detection experiments are reported in Table 2. Phrase boundaries occur after

Model	BOUNDARY $F_1$ (P,R)	Accuracy	Error
CRF	0.856 (0.839, 0.874)	92.15%	7.85%
BiRNN	0.867 (0.845, 0.896)	92.98%	7.02%

Table 2: *Intonational Phrase Boundary Detection results in Accuracy and F-measure for the BOUNDARY class.*

1,638 out of 5,900 words in the test data, resulting in a phrasing rate of 27.76%. Again, we report the  $F_1$  of the BOUNDARY class to more clearly describe the detection performance of these models. We find a relative reduction of error of **10.57%** on this task, along with an  $F_1$  improvement of 1.1% absolute. As in the accent detection experiments, we find consistent improvements to *both* precision and recall by the BiRNN model.

These results viewed together show that the BiRNN models these prosodic events quite well reducing errors by 10% compared the CRF, a very competitive, fairly well-optimized, baseline. Moreover, these improvements are not due to skewing the model to improve either precision or recall of the event. Rather, we observe consistent improvements to both.

## 7. Conclusion and Future Work

In this work we have demonstrated the value of sequential modeling for modeling prosodic events. In comparison to CRFs, Bidirectional Recurrent Neural Networks yield a relative reduction of error of approximately 10% in the recognition of both pitch accent (prominence) and intonational phrase boundaries. This improvement comes from two qualities of the BiRNN. First, the incorporation of both forward and backward history. Second, the amount of context encoded in the hidden, recurrent layers is learned from the data rather than being manually parameterized in the model configuration.

State-of-the-art performance at detecting these prosodic events is approaching human levels of performance. Future work will investigate the role of context and sequential models on distinguishing different pitch accent types and phrase ending tones (phrase accents and boundary tones), prosodic analysis tasks that remain difficult to automate. Moreover, it is very common, as we have done in this work, to treat the modeling of prominence and phrasing independently. Future efforts will investigate the interaction between these two prosodic events.

## 8. Acknowledgements

This work was partially funded by NSF IIS-1350550.

## 9. References

- [1] G.-A. Levow, "Context in multi-lingual tone and pitch accent recognition," in *Interspeech*, 2005.
- [2] S. Ananthakrishnan and S. Narayanan, "An automatic prosody recognizer using a coupled multi-stream acoustic model and a syntactic-prosodic language model," in *ICASSP*, 2005.
- [3] R. Fernandez and B. Ramabhadran, "Discriminative training and unsupervised adaptation for labeling prosodic events with limited training data," in *Interspeech*, 2010.
- [4] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, "ToBI: A standard for labeling english prosody," in *Proc. of the 1992 International Conference on Spoken Language Processing*, vol. 2, 1992, pp. 12–16.
- [5] A. Rosenberg, "Modeling intensity contours and the interaction between pitch and intensity to improve automatic prosodic event detection and classification," in *Interspeech*, 2012.
- [6] —, "AuToBI – a tool for automatic tobi annotation," in *Interspeech*, 2010.
- [7] —, "Automatic detection and classification of prosodic events," Ph.D. dissertation, Columbia University, 2009.
- [8] T. Mishra, V. R. Sridhar, and A. Conkie, "Word prominence detection using robust yet simple prosodic features," in *Interspeech*, 2012.
- [9] C. Wightman and M. Ostendorf, "Automatic labeling of prosodic patterns," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, October 1994.
- [10] M. Gregory and Y. Altun, "Using conditional random fields to predict pitch accents in conversational speech," in *ACL*, 2004.
- [11] A. Rosenberg, R. Fernandez, and B. Ramabhadran, "Phrase boundary assignment from text in multiple domains," in *INTER-SPEECH*, 2012.
- [12] R. Fernandez, A. Rendel, B. Ramabhadran, and R. Hoory, "Prosody contour prediction with long short-term memory, bi-directional, deep recurrent neural networks," in *INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association, Singapore, September 14-18, 2014*, 2014, pp. 2268–2272.
- [13] "The Stanford parser: A statistical parser," <http://nlp.stanford.edu/software/lex-parser.shtml>.
- [14] A. Rosenberg and J. Hirschberg, "Detecting pitch accents at the word, syllable and vowel level," in *HLT-NAACL*, 2009.
- [15] D. Talkin, "A robust algorithm for pitch tracking (rapt)," in *Speech Coding & Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds. Elsevier, 1995.
- [16] A. M. C. Sluijter, V. J. van Heuven, and J. J. A. Pacilly, "Spectral balance as a cue in the perception of linguistic stress," *Journal of the Acoustical Society of America*, vol. 101, no. 1, pp. 503–513, 1997.
- [17] K. Laskowski, J. Edlund, and M. Heldner, "An instantaneous vector representation of delta pitch for speaker-change prediction in conversational dialogue systems," in *ICASSP*, Las Vegas, NV, USA, 2008, pp. 5041–5044.
- [18] P. Taylor, "The tilt intonation model," in *ICSLP*, 1998.
- [19] N. Veilleux, J. Barnes, S. Shattuck-Hufnagel, and A. Brugos, "Perceptual robustness of the tonal center of gravity for contour classification," in *Workshop on Prosody and Meaning*, 2009.
- [20] R. Villing, J. Timoney, T. Ward, and J. Costello, "Automatic blind syllable segmentation for continuous speech," in *ISSC*, vol. 2004. IEEE, 2004, pp. 41–46.
- [21] A. Graves, *Supervised Sequence Labelling with Recurrent Neural Networks*. Springer, 2012.
- [22] F. A. Gers, J. Schmidhuber, and F. Cummings, "Learning to forget: Continual prediction with LSTM," *Neural Computation*, vol. 12, no. 10, pp. 2451–2471, 2000.
- [23] F. A. Gers, N. N. Schraudolph, and J. Schmidhuber, "Learning precise timing with LSTM Recurrent Networks," *J. of Machine Learning Research*, vol. 3, pp. 115–143, 2002.
- [24] A. Graves, "RNNLIB: A recurrent neural network library for sequence learning problems," <http://sourceforge.net/projects/rnnl/>.
- [25] C. Sutton and A. McCallum, "An introduction to conditional random fields for relational learning," in *Introduction to Statistical Relational Learning*, L. Getoor and B. Taskar, Eds. MIT Press, 2007.
- [26] T. Kudo. (2009) CRF++: Yet another CRF toolkit. [Online]. Available: <http://code.google.com/p/crfpp/>