Unsupervised Prominence Prediction for Speech Synthesis

Mahnoosh Mehrabani, Taniya Mishra, Alistair Conkie*

AT&T Labs Research
Florham Park, NJ, USA
mahmehrabani@utdallas.edu, {taniya,adc}@research.att.com

Abstract

We propose an unsupervised prominence prediction method for expressive speech synthesis. Prominence patterns are learned by statistical analysis of prosodic features extracted from speech data. The advantages of our unsupervised data-driven prominence prediction include: easy adaptation to new speakers, speech styles, and even languages without requiring expert knowledge or complicated linguistic rules.

In this approach, first, prominence predictive prosodic features are extracted at the foot level. Next, the extracted prosodic features are clustered, each cluster representing a prominence level. Based on just-noticeable-differences of prosodic features, the optimal number of perceptually distinct prominence levels is determined. Finally, the proposed prominence prediction is applied to prosody prediction for unit selection speech synthesis. Perceptual evaluation results show a preference for a 4-level unsupervised prominence prediction over a rule-based baseline in terms of naturalness and expressiveness of synthesized speech.
Index Terms: speech synthesis, prominence, prosody

1. Introduction

A basic aspect of human speech is *prosodic prominence*, which can be defined simply as what we hear when a word, syllable or a group of syllables is perceived as "standing out" from those around it in an utterance. Speakers use prominence to indicate the focus of an utterance, the introduction of new topics, the information status of a word (new or given), their emotion or attitude about the topic being discussed, or simply to draw the listener's attention. Predicting prominence as part of prosody prediction for speech synthesis has been shown to improve the naturalness and expressiveness of the generated synthetic speech [1, 2, 3, 4, 5].

A common approach for prominence prediction is to rely on hand-crafted rules to assign different degrees of prominence to select linguistic units (words, syllables, feet, etc). See [5, 6, 7, 1]. This approach though intuitive, introduces two issues. First, a rule-based approach typically requires a substantial investment in time and expert knowledge to be useful, and secondly will require extra work to adapt it to particular speaker or text-genre speech styles.

More recently data-driven approaches have been investigated. Given the relatively easy availability of spoken data, data-driven techniques offer the possibility of "learning" speaker and style specific prominence patterns from speech data without requiring detailed knowledge of linguistic theories for a particular language. Several authors have approached prominence prediction (or prominence annotation) in a data-driven way [6, 8, 9, 10, 11, 12, 13, 4]. In most of the examples cited, the classifiers for predicting prominence were built by training on data manually labeled with prominence tags. Due to the high cost of labeling in terms of time and money, a corpus of manually labeled data may be limited in size, making difficult to train classifiers with a high degree of precision. It can also be difficult to find labelers who can consistently and efficiently label prominence tags. Hence using a supervised data-driven model reduces the reliance on constructing rules for prominence prediction, but still requires significant resources to develop speaker or style-specific prominence prediction models.

We propose an unsupervised data-driven approach for prominence prediction. The approach relies on just-noticeable-differences in pitch, duration and energy to automatically estimate different degrees (or levels) of prominence from the speech corpus, and then models the relationship between different textual features and the prominence levels. Previously, an unsupervised approach for prominence annotation was taken by Wang and Narayanan [10] and Tambourini [9]. These works differ from ours in terms of underlying hypothesis, algorithm and end goal.

Prior work on prominence prediction and annotation have been performed at word, syllable, or syllable-nucleus level. In the approach described here, prominence is predicted at the foot level. A foot is defined as consisting of an accented syllable followed by all unaccented syllables that precede the next accented syllable or a phrase boundary. The use of feet as the unit for prominence prediction has been shown [14, 15] to be very effective in predicting local pitch contour shapes — which generally correlates with the perception of prominence.

2. Speech Database

For this study, we used a speech database consisting of approximately 50 hours of spoken material obtained from a single female speaker of American English. The speaker read a variety of textual material with good phonemic and prosodic coverage. The audio was recorded 16kHz, 16-bit in a studio environment. For the recorded material the speaker had a mean pitch of 219 Hz and standard deviation of 62 Hz. The speech material was labeled automatically with a number of features, including word, syllable and phoneme boundaries, based on forced alignment of text and audio. The data was also automatically annotated with additional information: syllable stress, phrase boundaries, and part of speech tags. Feet were marked automatically using syllable boundary, phrase boundary and syllable stress information. Roughly 500,000 feet were thus identified in the database from approximately 40,000 individual utterances. No manual labeling was performed on the database for our experiments.

* Mahnoosh Mehrabani is currently with the Center for Robust Speech Systems at the University of Texas at Dallas.

3. Unsupervised prominence prediction

Our proposed method for prominence prediction involves three main steps, described in Sections 3.1 – 3.3.

3.1. Prosodic feature extraction

For each foot, we developed an associated prosodic vector containing the following prosodic features, which were found to be the most predictive of perceived prominence in [16].

Energy-F0-Integral (EFI): This feature is intended to capture the increase in F0, duration and energy that is often associated with perception of prominence. It is the integral of the smoothed F0, energy and duration within the interval of the word, as shown in the equation below. (The variable α denotes a scaling factor that was set to 0.10).

$$EFI = \sum_{i \epsilon \text{interval}} (t_i \times F0_i \times \alpha \cdot \text{RMS-energy}_i) \qquad (1)$$

Voiced-to-Unvoiced Ratio (VUR): This feature was developed to act as *measure of reliability*. EFI is calculated on the smoothed F0. However, smoothing the F0 contour by interpolation over the unvoiced segments of the word may create spurious peaks or valleys in the F0 contour [17], and the resulting EFI may not reflect the "true" shape of the contour. VUR informs the model how much the EFI feature should be trusted. If the VUR is less than 0.5, then a majority of the segments in the word are unvoiced and most of F0 contour is obtained by smoothing, hence the EFI is less reliable than if the VUR was greater than 0.5.

Foot duration: The duration of the foot in number of 10 msec frames.

Aggregate statistics: This includes the mean, median, max, min, and variance of F0 and energy computed per foot.

These features were computed from pitch and energy contours extracted over 10 msec intervals for each utterance using ESPS Waves. Pitch halving and doubling were automatically cleaned up using an implementation of Bagshaw's 'defiltering' algorithm [18]. The F0 curve was smoothed over the unvoiced frames using weighted linear interpolation, where the weight vector was energy times voicing.

3.2. K-means clustering

Following development of the prosodic feature vectors, the feet in the database were clustered on the basis of their prosodic feature vectors using k-means clustering. K-means clustering partitions the data into a predefined number of clusters, defining k centroids: one for each cluster, and each data point is associated to the nearest centroid. The following objective function is minimized in order to calculate the centroids:

$$\sum_{i=1}^{k} \sum_{j=1}^{n} \|x_j - c_i\|^2 \qquad (2)$$

where $\|x_j - c_i\|$ is the chosen distance measure between a data point x_j and the cluster centroid c_i. In this study Euclidean distance was used for k-means clustering.

Based on the study described in [16] which showed that the listed prosodic features correlate with perception of prominence, we hypothesized that each cluster could represent a particular prominence level (or, a degree of prominence) associated with the feet in that cluster. Therefore, cluster centroids were examined next to determine the number of clusters, or the number of perceptually distinct prominence levels.

Feature	JND Unit Definition	Number of Clusters 2	3	4
Pitch	$\min(12 \times \log_2(\frac{F0_j}{F0_i}))$	2.5	1.6	1.3
Power	$\min(10 \times \log_{10}(\frac{e_j}{e_i}))$	2.1	1.5	1.1
Duration	$\min((d_j - d_i) \times 100/d_j)$	34.9	20.8	16.7

Table 1: *Minimum F_0 (in semitones), energy (in dB), and duration (% change) differences between cluster centroids.*

3.2.1. How many levels of prominence?

This is a significant question in prominence modeling. Literature review shows a variety of multi-level prominence scales were used in past studies. The most common is a binary scale indicating whether prominence is perceived or not. Other scales include a 3-level, 4-level, 11-level, and a 31-level scale, discussed in [19]. We are aware of three studies that have investigated the use of different scales [19, 20, 21], but with contradictory results regarding which is the best rating scale for prominence. In this work, we try to answer this question by focusing on our end goal: speech synthesis. We only want to model perceptually distinct prominence levels.

To identify the optimal number of perceptually distinct prominence levels, the centroids of the clusters of the foot-level prosodic feature vectors were examined using Just Noticeable Differences (JND) of pitch, duration, and energy. JND is the smallest perceivable difference between two levels of a sensory stimulus. Previous studies on JND for speech prosody [22, 23] suggest JND in pitch is 1.5 semitones, JND in energy is 0.5 dB, and JND in duration/tempo is a 10-15% change.

We calculated F_0, energy, and duration differences for every pair of cluster centroids, for systematic increase in the number of clusters. Table 1 summarizes the results for the number of clusters ranging from 2 to 4. The first, second, and third rows of the table represent minimum F_0, energy, and duration differences in semitones, dB, and percent, respectively, between cluster centroid i and j, such that $i \neq j$.

We hypothesized that the optimal number of prominence levels corresponds to the number of clusters, k, such that every pair of clusters is at least 1 JND apart in terms of pitch, duration and energy. We made the further assumption that if the smallest JND difference is greater than 2 JND, it is likely that the clusters are too far apart and certain perceptually distinct prominence levels are not being modeled, so prosodic vectors had to be reclustered into a greater number of clusters.

As shown in Table 1, when the number of prominence clusters was two, i.e., only two levels of prominence (prominent or not) was modeled, F_0, energy, and duration differences between the pair of clusters were substantially greater than 1 JND. For three prominence clusters, the cluster centroids were still more than 1 JND apart in terms of pitch, energy and duration. Increasing the number of clusters to four resulted in minimum pitch difference between a pair of clusters to be less than 1 JND (i.e, less than 1.5 semitones) but energy and duration differences

Prominence Level	0	1	2	3
2 Clusters	65.8%	34.2%		
3 Clusters	46.2%	40.1%	13.7%	
4 Clusters	34.8%	35.6%	22.8%	6.8%

Table 2: *Distribution of feet in 2, 3, and 4 prominence clusters.*

were still greater than 1 JND. Further increasing the number of clusters resulted in cluster differences substantially less than 1 JND for all three prosodic factors — and hence, by our assumptions unlikely to result in perceptually distinct prominence levels. Table 2 shows the distribution of prominence levels with two, three and four clusters. Prominence levels correspond to the magnitude of cluster centroids.

Based on the numbers in Table 1, we conjectured that prominence should be modeled either a 3-level or a 4-level scale. Though the minimum difference in pitch between a cluster pair for the 4-cluster condition was less than 1 JND, it is possible that the prominence levels corresponding that pair of (prosodic vector) clusters may still be perceptually distinct due to difference in duration and energy being greater than 1 JND. Whether the 3-level or the 4-level model is a better factorization of the prominence scale was evaluated through a perceptual evaluation (Sec. 4) of TTS output of a system in which prominence was modeled using our proposed approach. If the 4-levels of prominence that we have identified are indeed perceptually distinct, the 4-level model will produce more a lively and natural sounding synthetic speech than a 3-level prominence model.

3.3. Relating textual features to prominence

The final step of the proposed prominence prediction approach involved modeling the relationship between pertinent foot-level textual features and the perceptually distinct prominence levels identified through k-means clustering. The following textual features were considered:

- **T1:** Number of syllables in the foot.
- **T2:** The part-of-speech tag of the word containing the primary stressed syllable in the foot.
- **T3:** Information about whether the primary stressed syllable in the foot belongs to a function or content word.
- **T4:** Information about whether the primary stressed syllable's nucleus is a vowel, reduced vowel or a consonant.
- **T5:** Information about whether the primary stressed syllable belongs to a negation word or not.
- **T6:** The position of the foot in the phrase: phrase middle, phrase initial, or phrase final.

The textual features were selected on the basis of previous work that has shown correlations between these features and perception of prominence. Klabbers and van Santen [15] found significant correlation between number of syllables in a foot and the prominence-lending pitch movements. The correlation between part-of-speech tags and prominence has been noted in both early investigations [24], as well as newer works [13]. Prince and Smolensky [25] showed that the segmental makeup of a vowel nuclei had an effect on prosodic prominence. Wang and Narayanan [10] found that function word versus content word classification influenced prominence perception. Studies on the prosodic aspects of negation in English have shown that negative particles tend to be realized with pitch prominence [26]. And finally, the effect of position in the phrasal stream on prominence also has been demonstrated [27].

We built decision-tree-based models that predict prominence using the foot-level textual features as inputs. We built two prominence models: one that predicts 3-levels of prominence and another that predicts 4-levels. For the 3-level model, all textual features were significantly correlated (at $p < 0.001$) with the prominence tags. But for the 4-level model, T5 was not correlated with the prominence tags.

4. Prominence-based pitch prediction for text-to-speech synthesis

Our TTS engine is a unit-selection based engine. We incorporate prominence in our TTS system as an intermediate representation that is predicted from text and then used as the basis for computing the target pitch contour of the speech output. Our TTS system generates target pitch contours, in accordance with the General Superpositional Model of Intonation [28]. The pitch curve is a summation of a phrase curve and n accent curves. The phrase curve is created by interpolating over three points: the start of the phrase, the start of the last foot in the phrase, and the end of the phrase. The height of the phrase curve at the start of the phrase is the speaker's mean pitch, μ. The height of the phrase curve at the end of the phrase is $\mu - \alpha \cdot \sigma$, where σ is the standard deviation of the speaker's pitch, while α is a constant term that is determined empirically. $\alpha \cdot \sigma$ is never less than 2σ.

The accent curves are tied to the feet within a phrase. Each accent curve spans the length of a foot and is created by cosine interpolation over the start of the foot, the peak location within the foot, and the end of the foot. Peak location is a function of foot duration and the number of syllables in the foot, as determined in [15]. The height of the accent curve is determined by the prominence level predicted for the associated foot. The highest prominence level is assigned to the cluster whose centroid vector has the greatest magnitude, and the lowest prominence level is assigned to the cluster with the smallest magnitude centroid. The cluster corresponding to the lowest prominence level is considered to represent the feet that are not prominent. Accent curve heights range from 0 Hz (for non prominent feet) to $\beta \cdot 2\sigma$ Hz for the most prominence foot, where σ is the standard deviation of the speaker's pitch and β is the constant term that is empirically determined.

5. Perceptual evaluation

We conducted a perceptual experiment to evaluate the impact of the proposed unsupervised prominence prediction approach on the prosody of the synthetic speech output by our TTS engine compared to that of the rule-based prominence prediction approach, which serves as our baseline. This baseline system is based on the phonological rules described in [19], and has been empirically established to improve the prosody of TTS-generated speech over a pitch-prediction approach that is not based on the theory of prominence. Our perceptual evaluation was also designed to investigate the difference between the 3-level prominence rating scale versus the 4-level prominence rating scale on the prosody of the TTS-generated speech. These two rating scales were identified as most appropriate for our speech corpus considering the just-noticeable-differences in

pitch, duration and energy of the foot-based prosodic vectors in the corpus (Sec. 3.2.1).

5.1. Stimuli

The stimuli for the perceptual evaluation was generated from the text of twenty sentences. Ten of the twenty sentences were randomly selected from the MOCHATIMIT corpus [29]. And ten were randomly selected from sentences — derived from fiction — that contained the top 1000 most commonly used phrases in American English. The average sentence length was 13 words. Each sentence was synthesized by our TTS engine using three different prominence prediction models:

- UNSUP-3: Unsupervised 3-level prominence model.
- UNSUP-4: Unsupervised 4-level prominence model.
- BASELINE: Rule-based prominence model.

This resulted in a stimuli set of 60 utterances, in which each prominence model was represented by 20 stimuli utterances.

5.2. Listening protocol

The evaluation was conducted as a web-based listening test. The 60 stimuli utterances were presented in a single webpage. Each utterance was accompanied by the text of the utterance. The text was presented so that the listeners would focus more on the prosody rather than intelligibility of the synthetic sentences. The utterances could be played by clicking on a "play" button. The listeners were asked to pay attention to the prominence (explained as emphasis for non-expert listeners) patterns of each utterance and rate its overall quality (expressiveness and naturalness taken together) using the presented five-point Mean Opinion Score (MOS) scale, that ranged from Bad, Poor, Fair, Good, Excellent. The order of presentation of the 60 stimuli utterances were randomized for each listener. Listeners were also asked whether or not English was their native language, and whether they listened using headphones or speakers.

5.3. Results

24 listeners completed the task. Of the 24 listeners, 18 indicated that they were native speakers of English and 6 indicated that they were non-native. All listeners used headphones.

The results formed a 60 $\times$ 24 score matrix. Each column of the score matrix was z-transformed to eliminate differences in individual usage of the rating scales. From the z-transformed scores, we computed the mean score that each listener assigned to each model. A boxplot showing the distribution of mean scores for each of the three models is presented in Figure 1. Pairwise t-tests show that the differences between (i) UNSUP-4 and BASELINE are significant at $p < 0.0001$, t-value=4.4; and (ii) between UNSUP-3 and BASELINE are significant at $p < 0.0001$, t-value=2.4. But the difference between UNSUP-3 and UNSUP-4 is not statistically significant.

The mean of the z-normalized scores across all listeners (N = 24) for each model is this: UNSUP-3 = 0.005 $\pm$ 0.116 (implying average quality), UNSUP-4 = 0.088 $\pm$ 0.119 (implying above average quality), and BASELINE = -0.092 $\pm$ 0.114 (implying below average quality). 21 of the 24 listeners scored the utterances generated using one of the two unsupervised prominence prediction models higher than utterances generated by the rule-based model. Of these 21, 15 gave higher scores to utterances generated using UNSUP-4, while 6 gave higher scores to utterances generated using UNSUP-3.

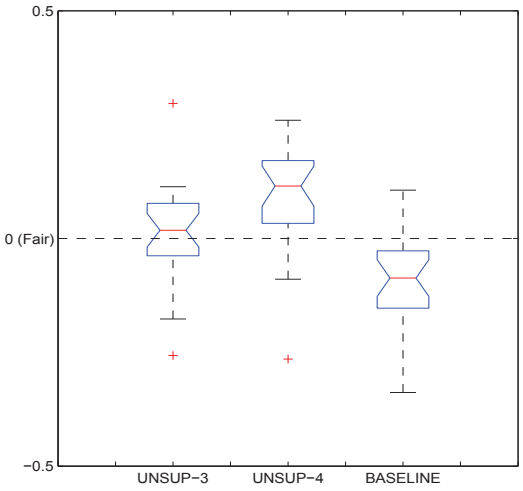

Figure 1: *Distribution of mean scores for each prominence prediction model.*

Overall, these results indicate that prosody generated on the basis of unsupervised data-driven prominence prediction models sounds better — considering naturalness and expressiveness — than prosody generated on the basis of prominence prediction rules. As for the question regarding whether the 3-level or the 4-level unsupervised prominence prediction model was more appropriate for the prosody prediction, the answer is less conclusive since the difference between the scores assigned to each of the two models was not statistically significant. The results however trend towards the 4-level prominence prediction.

6. Conclusions

An approach for "learning" the prominence patterns from speech data in an unsupervised manner for improved prosody prediction in speech synthesis systems is proposed. Our approach is based on the hypothesis that only perceptually distinct levels of prominence need to be modeled, as indicated by the Just-Noticeable-Differences in pitch, duration, and energy.

We extracted prominence predictive prosodic features at the foot-level from hours of speech data from a single speaker. The extracted feature vectors were clustered and cluster centroids were compared in terms of JND. We found that with four clusters (each cluster corresponding to a distinct level of prominence), the smallest difference between pairs of cluster centroids are close to one JND for pitch, duration, and energy. Increasing the number of clusters resulted in cluster centroid differences lower than one JND, which we hypothesized to be not perceivable and hence, unnecessary to model. The validity of our hypothesis was demonstrated through our perceptual experiment that showed that the 4-level data-driven prominence prediction model generated more expressive prosody in the synthesized speech compared to the rule-based baseline, which incidentally predicted 9-levels of prominence.

Overall our perceptual evaluation shows that our unsupervised prominence prediction approach is viable and our hypothesis is reasonable. In future studies, we will study how much the prominence prediction models generated by our approach generalize across speakers and styles. We will also work on incorporating data-driven prominence prediction scores directly in unit selection as search features.

7. References

[1] A. Windmann, I. Jauk, F. Tamburini, and P. Wagner, "Prominence-based prosody prediction for unit selection speech synthesis," in *Proc. Interspeech*, 2011.

[2] V. Strom, A. Nenkova, R. Clark, A. Vazquez-alvarez, J. Brenier, S. King, and D. Jurafsky, "Modelling prominence and emphasis improves unit-selection synthesis," in *in Proc. Interspeech*, 2007.

[3] J. Y. Zhang, A. R. Toth, K. Collins-Thompson, and A. W. Black, "Prominence prediction for super-sentential prosodic modeling based on a new database," in *IN 5TH ISCA SPEECH SYNTHESIS WORKSHOP*, 2004.

[4] C. W. Wightman, A. K. Syrdal, G. Stemmer, A. Conkie, and M. Beutnagel, "Perceptually based automatic prosody labeling and prosodically enriched unit selection improve concatenative text-to-speech synthesis," in *In Proceedings of the Intl. Conf. on Spoken Language Processing*, 2000, pp. 71–74.

[5] B. Heuft and T. Portele, "Synthesizing prosody: a prominence-based approach," in *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, vol. 3, October 1996, pp. 1361–1364 vol.3.

[6] C. Widera, T. Portele, and M. Wolters, "Prediction of word prominence," in *In Proc. European Conf. on Speech Communication and Technology*, 1997, pp. 999–1002.

[7] B. M. Streefkerk, L. C. Pols, and L. F. ten Bosch, "Acoustical and lexical/syntactic features to predict prominence," in *Proceedings of the Institute of Phonetic Sciences, University of Amsterdam*, vol. 24, 2001, pp. 155–165.

[8] M. O. D. Wong and J. Kahn, "Using weakly supervised learning to improve prosody labeling," University of Washington, Electrical Engineering Department, Seattle, Washington, Tech. Rep. UWEETR-2005-0003, January 2005. [Online]. Available: https://www.ee.washington.edu/techsite/papers/documents/UWEETR-2005-0003.pdf

[9] F. Tamburini, "Automatic prosodic prominence detection in speech using acoustic features: an unsupervised system," in *In Proceedings of Eurospeech 2003*, 2003, pp. 129–132.

[10] D. Wang and S. Narayanan, "An acoustic measure for word prominence in spontaneous speech," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 2, pp. 690–701, 2007.

[11] J. Yuan, J. M. Brenier, and D. Jurafsky, "Pitch accent prediction: Effects of genre and speaker," in *in Proc. Interspeech 2005*, 2005, pp. 1409–1412.

[12] N. Obin, X. Rodet, and A. Lacheret-Dujour, "French prominence: A probabilistic framework," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, 31 2008-April 4, pp. 3993–3996.

[13] V. Kumar, R. Sridhar, A. Nenkova, S. Narayanan, and D. Jurafsky, "Detecting prominence in conversational speech: pitch accent, givenness and focus," in *in Proc. Speech Prosody 2008*, 2008, pp. 453–456.

[14] E. Klabbers, J. Van Santen, and J. Wouters, "Prosodic factors for predicting local pitch shape," in *Proc. IEEE Workshop on Speech Synthesis*, 2002, pp. 123–126.

[15] E. Klabbers and J. Santen, "Clustering of foot-based pitch contours in expressive speech," in *Proc. Fifth ISCA Workshop on Speech Synthesis*, 2004.

[16] T. Mishra, V. R. Sridhar, and A. Conkie, "Word prominence detection using robust yet simple prosodic features," in *Proc. Interspeech*, 2012.

[17] T. Mishra, "Decomposition of fundamental frequency contours in the general superpositional intonation model," Ph.D. dissertation, Oregon Health and Science University, 2008. [Online]. Available: http://drl.ohsu.edu/cdm/ref/collection/etd/id/654

[18] P. C. Bagshaw, "Automatic prosodic analysis for computer aided pronunciation teaching," Ph.D. dissertation, University of Edinburgh, 1994.

[19] D. Arnold, P. Wagner, and B. Mbius, "Obtaining prominence judgments from nave listeners - influence of rating scales, linguistic levels and normalisation." in *INTERSPEECH*. ISCA, 2012.

[20] C. Jensen and J. Tndering, "Choosing a scale for measuring perceived prominence," in *INTERSPEECH 2005 - Eurospeech, 9th European Conference on Speech Communication and Technology, Lisbon, Portugal, September 4-8, 2005*. ISCA, 2005, pp. 2385–2388.

[21] C. Grover, B. Heuft, and B. V. Coile, "The reliability of labelling word prominence and prosodic boundary strength," in *ESCA and The University of Athens*, 1997, pp. 165–168.

[22] A. Rietveld and C. Gussenhoven, "On the relation between pitch excursion size and prominence," *Journal of Phonetics*, 1985.

[23] H. Quené, "What is the just noticeable difference for tempo in speech?" *On Speech and Language*, vol. 2, pp. 149–158, 2004.

[24] J. Hirschberg, "Pitch accent in context: Predicting intonational prominence from text," *Artificial Intelligence*, vol. 63, pp. 305–340, 1995.

[25] A. Prince and P. Smolensky, "Optimality theory: Constraint interaction in generative grammar," Rutgers Cognitive Science Center, Rutgers University, Tech. Rep., 1993.

[26] J. Hirschberg, "Accent and discourse context: assigning pitch accent in synthetic speech," in *Proceedings of the eighth National conference on Artificial intelligence - Volume 2*, ser. AAAI'90. AAAI Press, 1990, pp. 952–957. [Online]. Available: http://dl.acm.org/citation.cfm?id=1865609.1865642

[27] C. Bernard and J. Gervain, "Prosodic cues to word order: what level of representation?" *Front Psychol*, vol. 3, 2012.

[28] J. P. H. van Santen and B. Mbius, "A quantitative model of f0 generation and alignment," in *in IntonationAnalysis, Modelling and*, 2000, pp. 269–288.

[29] A. Wrench, "MOCHA-TIMIT," Department of Speech and Language Sciences, Queen Margaret University College, Edinburgh, Tech. Rep., 1999. [Online]. Available: http://www.cstr.ed.ac.uk/research/projects/artic/mocha.html