Automatic Detection of Sentence Prominence in Speech Using Predictability of Word-level Acoustic Features

Sofoklis Kakouros[1], *Okko Räsänen*[1]

[1] Department of Signal Processing and Acoustics, Aalto University, Finland

sofoklis.kakouros@aalto.fi, okko.rasanen@aalto.fi

Abstract

Automatic detection of prominence in speech is an important task for many spoken language applications. However, most previous approaches rely on the availability of a corpus that is annotated with prosodic labels in order to train classifiers, therefore lacking generality beyond high-resourced languages. In this paper, we propose an algorithm for the automatic detection of sentence prominence that does not require explicit prominence labels for training. The method is based on the finding that human perception of prominence correlates with the (un)predictability of prosodic trajectories. The proposed system takes speech as input and combines information from automatically detected syllabic nuclei and three prosodic features in order to provide estimates of the prominent words. Results are reported using a speech corpus with manually assigned prominence labels from twenty annotators, showing that the algorithmic output converges with the annotators' prominence responses with 86% accuracy.

Index Terms: Prominence detection, speech processing, speech analysis, syllabification, prominence perception

1. Introduction

Sentence prominence is an important property of speech where a speaker can convey meaning or intent that is not available in the linguistic message. In natural conversation, for instance, it is common that speakers make some words more prominent than others in order to draw the listener's attention to those parts of the utterance that carry the most information. In general, a speaker may use prosody in order to convey both linguistic and paralinguistic information. Respectively, listeners utilize the prosodic information in order to interpret speech. Methods capable of detecting prominence have therefore various uses in many spoken language applications such as automatic speech recognition (ASR). In contrast to traditional supervised approaches using manually annotated data (see, e.g., [1,2]), we propose a method that does not require training using annotated prosodic information and that is inspired by recent findings of human perception of prominence in speech. In the proposed scheme, two acoustic correlates of prominence are extracted from speech, modeled statistically over time, and combined with automatically detected syllabic nuclei information in order to provide estimates of the prominent words in the utterances.

Earlier work has suggested that prominence is a feature of speech that can attract human attention in a bottom-up manner [3]. This means that simply variations in the physical properties of the speech signal might be indicative of prominence. Such mechanism implies a rapid stimulus-driven response of the listener to the specific parts of the speech stream that are perceived to be more salient. Prosodic features, such as energy and F0, are good candidates for the representation of acoustic salience in speech as the speaker can modulate them relatively independently of the linguistic content. Bottom-up attention has been also defined in the literature as being a response to novel and unexpected (or unpredictable) stimuli [4]. In a recent study [5] it was shown that the temporal unpredictability of the fundamental frequency (F0) trajectories was connected with the perception of sentence prominence, thus giving support to the idea that unpredictability of the sensory stimulus is driving the listener's attention and thereby perception of prominence (see also [6] for an approach based on lexemes' prosodic features).

In this paper, we extend the earlier findings in [5] to a prominence detection system. We propose a method for the automatic detection of sentence prominence that does not require explicit prominence labels for training and that can capture prominent words in a manner hypothesized to be analogous to human perception.

1.1. Prosodic correlates of prominence

Prominence is a prosodic phenomenon that takes place at different domains in speech and can be generally described as an accentuation of syllables within words or of words within sentences [7]. The acoustic realization of prominence is typically manifested as changes in the fundamental frequency (F0), energy, and duration of the syllables or words [8]. Recent findings give also evidence of the importance of spectral tilt [9] as a correlate of prominence with, however, fewer studies supporting its role across languages [10]. Therefore, the present method focuses on the use of F0, energy, and duration.

1.2. Earlier work

When it comes to the automatic detection of prominence, there is extensive literature addressing the problem from various angles. The problem of automatic detection is particularly important as it can allow automatic tagging in large speech corpora and also allow machines to achieve a more naturalistic processing of speech. The major division in the research directions taken in the literature can be first split into approaches using linguistic information [11] (such as lexical or syntactic structure), acoustic information [12] (such as energy, F0, spectral tilt) or both [3,13]. The second domain of division is in supervised [14,15] and unsupervised techniques [3,16] where the major differentiating factor is on the usage of annotation data to train classifiers. Finally, the third and last division is based on the type of the statistical model utilized in order to train and evaluate the data [14,17]. Majority of the current approaches are typically supervised operating on both the linguistic and acoustic content of speech [1,2,14] whereas limited research has been done using unsupervised or semi-supervised methods [3,16]. For instance, Kalinli and Narayanan proposed a biologically inspired approach

combining bottom-up auditory attention cues together with high-level lexical and syntactic information achieving 83.11% accuracy using acoustic features only (unsupervised) and 85.71% using both acoustic and linguistic at the word level [3] (see also [18,19] for studies using clustering techniques).

In this study, we implement an approach using acoustic information, without prominence labels, in order to train a statistical n-gram model where we also make use of word duration information. The results are evaluated on an extensive set of annotated data that we collected and indicate that the proposed method is capable of detecting prominence with high agreement with the annotators' responses. Overall, our attempt is aimed at creating a method that can function similar to the human perception of prominence with performance reaching that of human labelers.

2. Methods

The proposed cognitively inspired algorithm for the automatic detection of sentence prominence (CADSP) consists of two main blocks (Fig. 1): (i) a method for the detection of syllabic nuclei and (ii) a statistical model that learns the typical prosodic trajectories in a set of training utterances and is then used to provide probability estimates of the prosodic trajectories in a set of novel utterances with unpredictable points corresponding to linguistic prominence (see [5]). These are further described below.

2.1. Signal envelope analysis for syllabic nuclei estimation

In order to estimate the number of syllabic nuclei in each word (or per time unit), signal amplitude envelope is used to segment speech into subsequent syllables. It is well known that the smoothed envelope of speech correlates with the syllabic structure of speech with the sonorant syllable nuclei corresponding to the envelope maxima while the syllabic boundaries roughly correspond to the minima between the nuclei (e.g., [20], [21]).

In order to compute the envelope, the absolute value of the speech signal sampled at 1000 Hz is first taken. The resulting signal is then low-pass filtered with a 48-ms moving average filter and scaled to have a maximum value of one across the signal length.

Syllabic boundaries are computed from the resulting envelope by simple minima detection: Any local minimum preceded by an amplitude larger than $\delta = 0.04$ is considered as a syllable boundary. Any two or more boundaries being closer than 80 ms to each other are considered as a one boundary at the midpoint of the boundaries. Finally, each local maximum in the envelope between the detected syllable boundaries is marked as a syllabic nucleus (see also [22] for a comparison with other methods).

2.2. Modeling prosodic trajectories

The aim of the algorithm is to mark words as prominent if the temporal evolution of the prosodic features is unpredictable during the words, violating the expectations of the listener (or the model) and thereby capturing the attention.

Earlier research has indicated F0, energy, and duration as being the three features in speech that are most descriptive of prominence across a number of languages (see, e.g., [8,10,23,24]). In the current work, F0 and energy are used to train a statistical n-gram model that describes the typical temporal evolution of these features on a set of training utterances. This model is then used to provide expectations

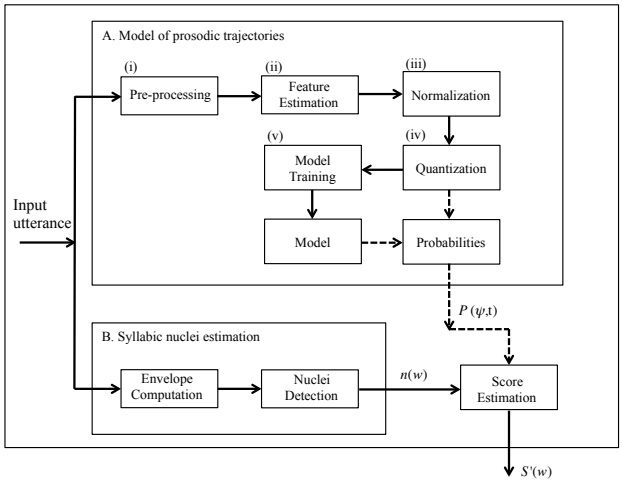

Figure 1: *Overview of the processing steps in the algorithm.*

(probabilities) for the prosodic features across time on novel input. The effect of duration is included in the energy and F0 features through integrating the probabilities of these features over word duration.

The model consists of five processing steps (see Fig. 1): (i) pre-processing, (ii) feature extraction, (iii) feature normalization, (iv) quantization, and (v) n-gram parameter computation (during training) / probability estimation (during testing).

As a pre-processing step, the speech data are downsampled to 8 kHz. The two prosodic features are then computed: F0 contours for the voiced segments are extracted from each utterance using the YAAPT algorithm [25] with a 25-ms window and 10-ms step size while energy is computed using the same window length and step size as follows:

$$E = \sum_{n=n_1}^{n_2} |x[n]|^2 , \quad (1)$$

where x is the speech input and n_1, n_2 define the beginning and end of the analysis window respectively.

In order to ensure comparability of the features across utterances and talkers, F0 and energy are min-max normalized according to Eq. (2).

$$f'(t) = \frac{f(t) - \min(f)}{\max(f) - \min(f)} \quad (2)$$

In the equation, f denotes the feature value at time t while max(f) and min(f) refer to the maximum and minimum values of the feature, respectively, during the given utterance (see [1]).

The extracted features for each time frame are quantized into Q discrete amplitude levels, $f'(t) \rightarrow a_t \in [1, 2, \ldots, Q]$, in order to allow discrete probability modeling of the data. In the present study, $Q = 32$ quantization levels were computed using the k-means algorithm with a random sample initialization. The number of levels was selected as a compromise between the best approximation of the feature contours and the least number of discrete levels possible.

For the statistical modeling of the temporal evolution of the feature trajectories, n-gram probabilities are computed from the relative frequencies of different n-tuples in the training data:

$$P_\psi(a_t \mid a_{t-1}, \ldots, a_{t-n+1}) = \frac{C_\psi(a_t, a_{t-1}, \ldots, a_{t-n+1})}{C_\psi(a_{t-1}, \ldots, a_{t-n+1})} , \quad (3)$$

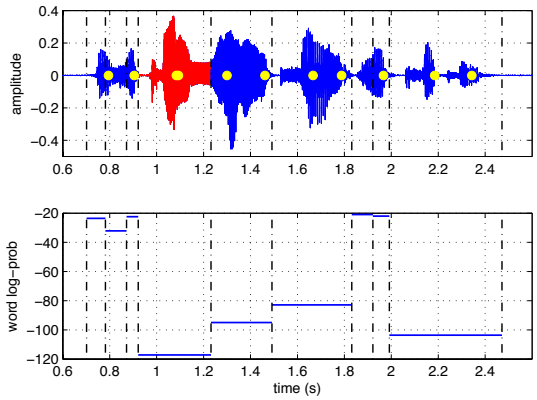

Figure 2: *Example output of the algorithm for the utterance "There is a clean yellow cow and a cookie". Top panel: original signal waveform where red marks the word perceived as prominent by the majority of the listeners and yellow marks the syllabic nuclei. Bottom panel: word scores produced by the proposed algorithm.*

where C denotes the frequency counts of the discrete n-tuples and ψ the feature in question. The probability $P'(t)$ of the features at time t is computed according to Eq. (4), i.e., by summing the log-probabilities over the two features ψ of interest. This formulation assumes that the features are independent of each other.

$$P'(t) = \sum_{\psi} \log\left(P_{\psi}(a_t \mid a_{t-1}, \ldots, a_{t-n+1})\right) \quad (4)$$

Standard n-grams were chosen due to their relative simplicity and scalability. Analysis is limited to n-gram orders of n = 2, 3, and 4. Bi-grams (n = 2) represent the shortest temporally ordered segment while four-grams (n = 4) are the longest sequence for which probabilities can be reliably estimated from the dataset used in the experiments.

In order to measure the overall predictability of the prosody during each word, F0 and energy-based word-level prominence scores $S(w_{i,j})$ are computed for each word $w_{i,j}$ in utterance i by integrating the instantaneous feature probabilities over the duration of the entire word:

$$S(w) = \sum_{t=t_1}^{t_2} P'(t)\,. \quad (5)$$

Here, temporal boundaries, t_1 and t_2, are extracted from the word-level transcription of the speech database but could be also obtained automatically, from, e.g., an ASR system.

In order to use a subjective measure of the syllable duration, the average duration of the syllables within a word are computed by dividing the word duration t_w by the number of nuclei v detected for that word ($t_v = t_w/v$). As duration is one important correlate of prominence in speech, each acoustic feature based word score in the utterance is weighted by the exponent of the average syllable duration (see Fig. 2). Longer syllabic durations lead typically to increased perception of prominence and therefore the exponential function is good for the non-linear mapping of the durational nucleic information.

$$S'(w) = S(w) \cdot e^{t_v} \quad (6)$$

The prominence classification $H(w_{i,j})$ for each word j in utterance i is then determined based on whether the word-level score $S'(w_{i,j})$ falls below a threshold r_i:

$$H(w_{ij}) = \begin{cases} 1, & S'(w_{ij}) < r_i\,, \\ 0, & S'(w_{ij}) \geq r_i\,, \end{cases} \quad (7)$$

where the threshold is defined at the utterance level as

$$r_i = \mu_i - \sigma_i \lambda \quad (8)$$

and where hyperparameter λ controls the sensitivity of the prominence detector.

3. Experiments

The performance of CADSP was tested on continuous English speech. In order to evaluate algorithmic output, 20 naïve listeners were invited to mark prominence. Their annotations were compared against the prominence hypotheses generated by the CADSP algorithm. Overall performance was evaluated using standard measures for accuracy and inter-rater agreement that are further described below.

3.1. Material

The CAREGIVER Y2 UK corpus [26] was used in the experiments. The style of speech in CAREGIVER is acted infant-directed speech (IDS) spoken in continuous UK English and recorded in high quality within a noise-free anechoic room. The talkers were not separately instructed on the use of prosody or prominence (see [26], for details). In overall, the CAREGIVER Y2 UK corpus contains speech from 10 adult talkers with a total duration of approximately 8.7 hours. Specifically, the corpus contains data from 4 primary talkers (2 male, 2 female) producing 2397 utterances each and 6 secondary talkers (3 male, 3 female) producing 600 utterances each. A subset of 300 unique utterances was chosen for the listening tests from one male and one female primary talker (Speakers 3 and 4), yielding a total of 600 sentences for evaluation (test set). All single-word sentences were excluded from the data and there were 5.9 words per sentence on average. The corpus was chosen for this study as there are extensive speech data available per speaker that is beneficial for the training and evaluation purposes (e.g. testing different n-gram orders) and was also readily available to us. In addition, the prominence markings collected for the corpus have agreement rates equivalent to those of studies in adult-directed American English (see [27] for more details).

3.2. Data collection

A total of twenty subjects (11 male, 9 female, age range 20-61 with a median of 30 years) participated in a listening experiment where they were asked to mark perceived prominence in the test set. The participants were recruited among the students and personnel of Aalto University and University of Helsinki, Finland. The majority of the participants (N=14) were L1 (first language) Finnish speakers, while the remaining (N=6) were L1 UK English speakers. English was the L2 (second language) of all Finnish listeners and each Finnish listener in the experiment reported to be a professional-level English speaker (see also [27]). Six of the L1 Finnish subjects also took the LexTALE proficiency test on English as a post-hoc control procedure, achieving an average score of 92.08/100 in the test, corresponding approximately to C1 & C2 in the Common European Framework (CEF) language proficiency levels [28]. No significant differences were observed in the coherence of the annotations between the two L1 groups [29]. All participants reported normal hearing.

3.3. Evaluation

Precision (PRC), recall (RCL), their harmonic mean (F-value), and accuracy (ACC) were used as the primary quality measures and were defined as:

$$\mathrm{RCL} = tp / (tp + fn) \quad (9)$$

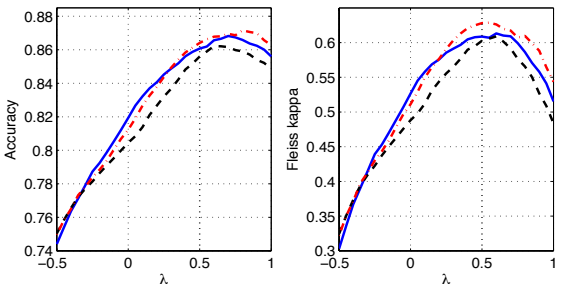

Figure 3: *Prominence detection performance. Blue sold line marks F0, black dashed line marks energy and red dashed-dotted line marks the combined performance of energy with F0.*

$$\mathrm{PRC} = tp / (tp + fp) \quad (10)$$

$$\mathrm{F} = (2 \times \mathrm{PRC} \times \mathrm{RCL})/(\mathrm{PRC} + \mathrm{RCL}) \quad (11)$$

$$\mathrm{ACC} = (tp + tn)/(tp + fp + fn + tn) \quad (12)$$

where *tp* denotes the true positives, *tn* the true negatives, *fp* the false positives, and *fn* the false negatives.

Standard Fleiss kappa [30] was used in order to measure the pairwise agreement rates between the algorithm and the annotators. Also, it allows comparison of our results with that of other similar studies in prominence detection (see, e.g., [31,32]). Overall, Fleiss kappa measures the degree of agreement between two or more annotators on a nominal scale of $\kappa \in [-1,1]$ and yields $\kappa = 0$ if the number of agreements is equal to what is expected based on chance-level co-occurrences in the data and $\kappa = 1$ if all annotators fully agree.

In this work, Fleiss kappa was measured on the word-level. For each word occurring in the test set, a binary decision between non-prominence and prominence was considered. From the set of twenty annotators, an annotation reference was generated based on majority agreement where 22.8% of all words were prominent. The overall agreement rate on the prominence markings of words in the reference set was then used as the primary performance measure.

3.4. Results

The experiment was run in a cross-validation setup where data from 9 speakers were used for training and one for testing. Three orders of the n-gram model (n = 2, 3, and 4) were used for training on a set of 9594 utterances from 9 speakers (≈7.2 hours of data) and testing on the held-out set of 300 annotated utterances (≈30 minutes of data) on one of the two annotated speakers, leading to two runs of the experiment. None of the test signals were used in training. To compare against a baseline, a random reference was also generated where h words were randomly marked as prominent in each utterance in the test set, where h denotes the number of hypotheses generated by the algorithm. The reference was run over 10 iterations and the results show that random selection of prominent words gives $\kappa = 0.01$ ($\sigma = 0.01$) indicating no agreement.

In terms of the individual features' performance, F0 (ACC=86.20%) and energy (ACC=86.16%) seem to be equally descriptive in determining prominence (see Table 1). Syllable duration alone has much lower F and kappa measures, indicating that its function independently of the other features does not explain prominence as accurately. Several feature combinations were also tested and the best performance was achieved for energy and F0 (ACC=86.95%). N-gram order does not have a large effect on the results and therefore the results in the table are pooled across the n-gram orders

Table 1. *Prominent word detection performance for the individual features and their combination (for λ = 0.7) pooled over the three n-gram orders (n = 2, 3 and 4) and averaged across speakers.*

	ACC	F	PRC	RCL	Fleiss Kappa
F0+EN	86.95% ±0.18	71.99% ±0.39	73.25% ±0.13	70.78% ±0.87	0.61 ±0.04
F0	86.20% ±0.39	71.22% ±0.99	73.60% ±1.01	69.00% ±0.97	0.60 ±0.01
EN	86.16% ±0.34	70.15% ±0.57	71.02% ±1.21	69.31% ±0.05	0.59 ±0.09
Syllable duration	80.72% ±0.15	53.90% ±0.10	56.85% ±0.17	51.23% ±0.10	0.37 ±0.01

(σ <0.005 between n-gram orders). It is suggested to use low-order n-grams if there is only little speech data available for the model training in order to enable reliable estimation of the probabilities. For high order n-grams (> 4), the performance deteriorates as the training data becomes too sparse even for a large corpus.

Figure 3 shows the performance of the algorithm as a function of the detection threshold λ. It can be seen that for λ > 0.5 the algorithm converges to its highest performance reaching an accuracy of 86.95% and Fleiss kappa of 0.61. This level of performance compares well with other approaches that do not use prosodic labels (see, e.g., [3,11,12,14,17]). For instance, Kalinli and Narayanan [3] report accuracy of 83.11% while Chen et al. achieved 77.30% [13], both on BU-RNC database [33] using only acoustic features on word level. Similarly Tamburini and Caini [12] achieved 80.6% on TIMIT database using acoustic features on syllable level. Direct comparison of the results is not possible due to the use of different speech corpora. Nonetheless, the results (BU-RNC and CAREGIVER) may be at least partly comparable as they are described by similar inter-annotator agreement rates (see [27] and [33] for more information). As this study was conducted on a specific speaking style, we can currently only conclude that the proposed approach seems to model and perform well on IDS. Future work will include evaluation on more speaking styles and languages.

4. Conclusions

A computationally simple method for the automatic detection of sentence level prominence was presented in this paper. The aim was to build a system that functions in a manner analogous to the hypothesized human prominence perception mechanism presented in [5]. We proposed to use the cumulative word-level unpredictability of the most important acoustic correlates of prominence coupled with a syllabic-nuclei-based weighting scheme in order to detect prominent words. Performance of the system was tested by comparing the algorithmic output with that of the perception of prominence by human listeners and evaluated in a number of different experiments. The model showed accuracy of 86% with the annotators' responses providing initial promising results for the method. In future work, the algorithm needs to be further evaluated with other types of speech and in more languages in order to ensure generalizability.

5. Acknowledgements

This research was performed as a part of the Data to Intelligence (D2I) project funded by Tekes, Finland and by the Academy of Finland in the project "Computational modeling of language acquisition".

6. References

[1] K. Imoto, Y. Tsubota, A. Raux, T. Kawahara, and M. Dantsuji, "Modeling and Automatic Detection of English Sentence Stress for Computer-Assisted English Prosody Learning System," in *Proceedings of the Seventh International Conference on Spoken Language Processing*, pp. 749–752, 2002.

[2] N. Minematsu, S. Kobashikawa, K. Hirose, and D. Erickson, "Acoustic Modeling of Sentence Stress Using Differential Features Between Syllables for English Rhythm Learning System Development," in *Proceedings of the Seventh International Conference on Spoken Language Processing (ICSLP'2002)*, pp. 745–748, 2002.

[3] O. Kalinli and S. Narayanan, "Prominence detection using auditory attention cues and task-dependent high level information," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 5, pp. 1009–1024, 2009.

[4] L. Itti and P. Baldi, "Bayesian surprise attracts human attention," *Vision Research*, vol. 49, pp. 1295–1306, 2009.

[5] S. Kakouros and O. Räsänen, "Statistical unpredictability of F0 trajectories as a cue to sentence stress," in *Proceedings of the 36th Annual Conference of the Cognitive Science Society,* pp. 1246–1251, Quebec, Canada, 2014.

[6] S. Kakouros and O. Räsänen, "Analyzing the Predictability of Lexeme-specific Prosodic Features as a Cue to Sentence Prominence," in *Proceedings of the 37th Annual Conference of the Cognitive Science Society,* Pasadena, California, 2015.

[7] S. Werner and E. Keller, "Prosodic aspects of speech", in Keller, E. (Ed.), *Fundamentals of Speech Synthesis and Speech Recognition: Basic Concepts, State of the Art, and Future Challenges*. Chichester: John Wiley, pp. 23–40, 1994.

[8] P. Lieberman, "Some acoustic correlates of word stress in American English," *J. Acoust. Soc. Am.*, vol. 32, no. 4, pp. 451–454, 1960.

[9] A. M. C. Sluijter and V. J. van Heuven, "Spectral balance as an acoustic correlate of linguistic stress," *J. Acoust. Soc. Am.*, vol. 100, no. 4, pp. 2471–2485, 1996.

[10] M. Ortega-Llebaria and P. Prieto, "Acoustic correlates of stress in central Catalan and Castilian Spanish," *Language and Speech*, vol. 54, no. 1, pp. 1–25, 2010.

[11] S. Pan and J. Hirschberg, "Modeling local context for pitch accent prediction," in *Proceedings of the 38th annual meeting on association for computational linguistics*, pp. 233–240, 2000.

[12] F. Tamburini and C. Caini, "An automatic system for detecting prosodic prominence in American English continuous speech," *International Journal of Speech Technology*, vol. 8, pp. 33–44, 2005.

[13] K. Chen, M. Hasegawa-Johnson, A. Cohen, and J. Cole, "A maximum likelihood prosody recognizer," in *Proceeding of Speech Prosody*, Nara, Japan, pp. 509–512, 2004.

[14] G. Christodoulides and M. Avanzi, "An evaluation of machine learning methods for prominence detection in French," in *Proceedings of the Fifteenth Annual Conference of the International Speech Communication Association (Interspeech)*, pp. 116–119, 2014.

[15] H. Moniz, A. I. Mata, J. Hirschberg, F. Batista, A. Rosenberg, and I. Trancoso, "Extending AuToBI to prominence detection in European Portoguese," in *Proceedings of Speech Prosody*, 2014.

[16] J. H. Jeon and Y. Liu, "Automatic prosodic event detection using a novel labeling and selection method in co-training," *Speech Communication*, vol. 54, pp. 445–458, 2012.

[17] F. Tamburini, C. Bertini, and M. P. Bertinetto, "Prosodic prominence detection in Italian continuous speech using probabilistic graphical models," in *Proceedings of Speech Prosody*, pp. 285–289, 2014.

[18] S. Ananthakrishnan, and S. Narayanan, "Combining acoustic, lexical, and syntactic evidence for automatic unsupervised prosody labeling," in *Proceedings of Interspeech*, pp. 297–300, 2006.

[19] G. A. Levow, "Unsupervised and semi-supervised learning of tone and pitch accent," in *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pp. 224–231, 2006.

[20] P. Mermelstein, "Automatic segmentation of speech into syllabic units," *J. Acoust. Soc. Am.,* vol. 58, no. 4, pp. 880–883, 1975.

[21] R. Villing, T. Ward, and J. Timoney, "Performance limits for envelope based automatic syllable segmentation," in *Proceedings of ISSC'2006*, Dublin, Ireland, June 28–30, 2006.

[22] O. Räsänen, G. Doyle, and M. C. Frank, "Unsupervised word discovery from speech using automatic segmentation into syllable-like units," in *Proceedings of Interspeech*, Dresden, Germany, 2015.

[23] J. Terken, "Fundamental frequency and perceived prominence of accented syllables," *J. Acoust. Soc. Am.*, vol. 89, no. 4, pp. 1768–1776, 1991.

[24] G. Kochanski, E. Grabe, J. Coleman, and B. Rosner, "Loudness predicts prominence: Fundamental frequency lends little," *J. Acoust. Soc. Am.*, vol. 118, no. 2, pp. 1038–1054, 2005.

[25] S. A. Zahorian and H. Hu, "A spectral/temporal method for robust fundamental frequency tracking," *J. Acoust. Soc. Am.*, vol. 123, no. 6, pp. 4559–4571, 2008.

[26] T. Altosaar, L. ten Bosch, G. Aimetti, C. Koniaris, K. Demuynck, and H. van den Heuvel, "A speech corpus for modeling language acquisition: CAREGIVER", in *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pp. 1062–1068, 2010.

[27] S. Kakouros and O. Räsänen, "Perception of sentence stress in English infant directed speech," in *Proceedings of Interspeech,* Singapore, 2014.

[28] K. Lemhöfer and M. Broersma, "Introducing LexTALE: A quick and valid Lexical Test for Advanced Learners of English," *Behavior Research Methods,* vol. 44(2), pp. 325–343, 2012.

[29] S. Kakouros and O. Räsänen, "Perception of sentence stress in speech correlates with the temporal unpredictability of prosodic features," submitted.

[30] J. L. Fleiss, "Measuring norminal scale agreement among many raters", *Psychological Bulletin,* vol. 76, pp. 378–382, 1971.

[31] Y. Mo, J. Cole, and E-K. Lee, "Naïve listeners' prominence and boundary perception," in *Proceedings of Speech Prosody*, Campinas, Brazil, pp. 735–738, 2008.

[32] H-J. You, "Determining prominence and prosodic boundaries in Korean by non-expert rapid prosody transcription," in *Proceedings of Speech Prosody*, Shanghai, China, 2012.

[33] M. Ostendorf, P. Price, and S. Shattuck-Hufnagel, "The Boston University radio news corpus," *Linguistic Data Consortium*, pp. 1–19, 1995.