Analyzing the Predictability of Lexeme-specific Prosodic Features as a Cue to Sentence Prominence

Sofoklis Kakouros (sofoklis.kakouros@aalto.fi)
Department of Signal Processing and Acoustics, Aalto University,
PO Box 13000, AALTO, Finland

Okko Räsänen (okko.rasanen@aalto.fi)
Department of Signal Processing and Acoustics, Aalto University,
PO Box 13000, AALTO, Finland

Abstract

This study investigates the relationship between sentence prominence and the predictability of word-specific statistical descriptors of prosody. We extend from an earlier word-invariant model by studying a model that marks words as prominent if the acoustic prosodic features differ from their expected values during the lexemes. To test the approach, the most common acoustic features associated with the perception of prominence are extracted and several lexeme-specific statistical measures are computed for each feature. Simulations are conducted on a corpus of continuous English speech and the algorithm output is compared to manually assigned prominence labels. The results show that the deviant prosodic descriptors of the words correlate with the perception of prominence. However, this effect is much smaller than that obtained by modeling the prosodic predictability at the utterance level, suggesting that context-independent lexeme-specific models are unable to capture relevant aspects of sentence prominence.

Keywords: Sentence prominence; prosody; statistical learning; predictability; attention

Introduction

Sentence prominence or stress is an important characteristic of spoken language that can be defined as an accentuation of syllables within words or of words within sentences (Cutler, Dahan, & van Donselaar, 1997). Prominence has an impact on the perceptual processing of the listener, where, however, little is known about the actual mechanism or process that drives the perception of prominence. This study extends our earlier work where we examined how the temporal unpredictability of F0 affects the perception of prominence and was agnostic to the lexical content of the utterances (Kakouros & Räsänen, 2014a). Here, the aim is to investigate whether there are interactions between the lexical and prosodic coding of each word through the investigation of the most common acoustic correlates of prominence that occur during the words.

The perception of prominence is largely determined by contrastive changes in prosodic features estimated over temporally defined segments such as those of a word, sentence or of longer utterances (Werner & Keller, 1994). Recent studies have also associated prominence with the function of attention as the mechanism enabling the shift of focus to specific words in an utterance (see, e.g., Cole, Yo, & Hasegawa-Johnson, 2010; Kalinli & Narayanan, 2009). For instance, Cole et al. (2010) concluded that attention and prominence might share a common basis where a word may attract the listener's attention either as a response to acoustic modulation (signal-based acoustic salience) or due to its relative unpredictability requiring extra processing resources (expectation-based). Therefore, in this regard, attention can be roughly divided into a bottom-up and top-down component (see also Mancas, Beul, Riche, & Siebert, 2012).

Bottom-up is a rapid, saliency-driven component while top-down is a task-dependent process that involves high-level cognitive processes (see, e.g., Mancas et al., 2012, for more details) and is considered to use prior knowledge and past expertise (see also Kalinli & Narayanan, 2009). Both attentional components are assumed to interact and, according to Itti and Baldi (2009), one way to characterize attention is by the unexpectedness or novelty of stimulus that can be converted into a probabilistic interpretation under a statistical learning framework. One such formulation is that of a low likelihood data observation taking place in an otherwise predictable temporally defined context. For instance, assuming a series of data observations O_t during $[t_1, t_N]$, with $P(O_t)$ their corresponding likelihood, the observation O_m that would provide the lowest probability given the past learned expectations (O_m=min{$P(O_{t1})$,$P(O_{t2})$,…,$P(O_{tN})$}) would be the one characterized as novel. This can be extended to identifying multiple novel observations by selecting local minima or through the use of a probabilistic threshold. An analogy to speech can be found in the prosodic features where the acoustic information can be statistically modeled and evaluated over different temporal segments such as those of individual words, therefore combining top-down (lexical) and bottom-up (prosodic) processing.

The relationship between language, acoustic features of speech, and predictability has been also examined earlier. For instance, works such as that of Calhoun (2007, 2010) point to the importance of predictability of the linguistic elements (such as syntactic and semantic) in determining focus in speech. Cole et al. (2010) also examined the role of expectations, reporting the importance of word unpredictability in the perception of prosodic prominence. Finally, another related work is that of Aylett and Turk's (2004) smooth signal redundancy hypothesis. The hypothesis is based on the relationship between syllable reduction (through durational shortening) and linguistic predictability and proposes that prosodic prominence is employed in order to manage unpredictable elements in speech (see also Turk, 2010, for a similar study on words; Pan & Hirschberg, 2000, for a study using only the lexical context; Aylett & Bull, 1998). Therefore studying the potential interactions between the acoustic and language content is particularly important.

The acoustic realization of prosodic prominence is typically associated with the acoustic features corresponding to signal energy, fundamental frequency (F0), and duration (see, e.g., Lieberman, 1960; Terken, 1991; Kochanski, Grabe, Coleman, & Rosner, 2005; Wagner, 2005; Rosenberg & Hirschberg, 2009). Few studies give evidence of the importance of spectral tilt (see, e.g., Sluijter & van Heuven, 1996) with, however, experimental findings, not being able, thus far, to confirm its role across languages (Ortega-Llebaria & Prieto, 2010). When it comes to computational modeling of sentence prominence, the majority of the earlier work has focused on supervised (see, e.g., Imoto, Tsubota, Raux, Kawahara, & Dantsuji, 2002; Minematsu, Kobashikawa, Hirose, & Erickson, 2002) and unsupervised approaches (see, e.g., Kalinli & Narayanan, 2009; Wang & Narayanan, 2007; Tamburini & Caini, 2005). Supervised approaches have been primarily studied by examining the co-occurrence statistics of combinations of the typical prosodic features and the perception of prominence (see e.g., Imoto et al., 2002; Minematsu et al., 2002). This typically requires the availability of manually annotated prominence labels, which is an overall expensive process. Instead of using a priori linguistic information, unsupervised methods typically extract acoustic features directly from the speech signal and compute, for instance, prominence scores using different feature combinations (see, e.g., Tamburini & Caini, 2005).

In Kakouros & Räsänen (2014a) it was proposed that the unpredictability of temporally evolving prosodic features could be sufficient for generating a perception of prominence in speech, therefore making prominence perception learnable using generic statistical learning mechanisms. In the current paper, we investigate the idea of predictability-based prominence perception further by combining lexical information to the unpredictability framework. The basic assumption is that typical (high probability) prosodic feature values during a specific lexeme would correspond to a non-prominent word whereas deviant (low probability) values would be surprising to the listener and therefore constitute a potentially prominent word.

Methods

The vocabulary-based acoustic parameters (VAP) approach is centered on a dictionary of words where the prosodic characteristics of each word are described with a number of statistical descriptors (see Fig. 1). Although unrealistic for large-vocabulary languages, the approach enables a controlled way to study the expectations of prosodic features using data that has multiple occurrences of each word. Each descriptor is modeled using a parametric and a non-parametric distribution in order to infer typicality of the word realization in the data, and thereby to detect potentially prominent words based on their atypical prosodic characteristics. These are further described below, starting with the prosodic features estimation.

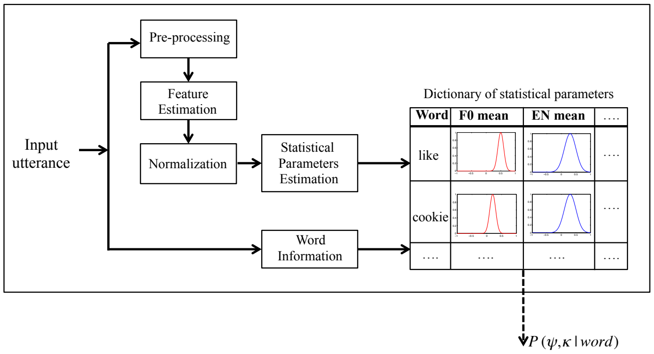

Figure 1: Overview of the processing steps during training and testing for the proposed algorithm.

Features and their statistical descriptors

Four features descriptive of prominence were used in the data analysis, namely: (i) signal energy (EN), (ii) F0, (iii) spectral tilt (ST), and (iv) duration (D).

Energy, F0, and spectral tilt The speech data were first downsampled to 8 kHz. For the voiced segments, F0 contours were extracted for each utterance using the YAAPT-algorithm (Zahorian & Hu, 2008) with 25 ms window length and 10 ms frame shift. Signal energy was calculated using the same window size and frame shift based on Eq. (1):

$$EN = \sum_{n=n_1}^{n_2} |x[n]|^2 \tag{1}$$

Spectral tilt was computed using the same windowing parameters and by taking the first Mel-frequency cepstral coefficient (MFCC) of each window (see, e.g., Tsiakoulis, Potamianos, & Dimitriadis, 2010).

Before calculating the statistical descriptors for each feature, a normalization process was applied in order to ensure comparability across talkers and utterances. Energy and spectral tilt were min-max normalized per utterance according to Eq. (2):

$$f_\psi'(t) = \frac{f_\psi(t) - \min(f_\psi)}{\max(f_\psi) - \min(f_\psi)} \tag{2}$$

where $f_\psi(t)$ represents the value of feature ψ at time t and the min and max are computed across the entire utterance (see also Imoto et al., 2002). F0 contours were semitone normalized based on the minimum F0 during the utterance, according to Eq. (3):

$$F0'(t) = 12 \cdot \log_2 \frac{F0(t)}{\min(F0)} \tag{3}$$

where *F0(t)* represents the value of the F0 at time *t*. Finally, word- and syllable-level durations were modeled in their original form.

Word- and syllable-level duration Duration in the data was examined both at the word and syllable level. Word duration (D_w) was computed from the time-aligned word-level information extracted from the transcriptions of the speech corpus. Specifically, given the temporal boundaries of each word in the utterances, t_1 (word start) and t_2 (word end), duration was calculated as $D_w = t_2 - t_1$.

Syllable duration (D_s) was computed by dividing the word duration D_w by the number of syllabic nuclei v detected for that word ($D_s = D_w / v$). In order to estimate the number of syllabic nuclei in each word (or per time unit), the amplitude envelope was used to segment the speech signal into subsequent syllables. For the envelope computation, the absolute value of the speech signal sampled at 1000 Hz was taken first followed by a low-pass filtering with a 48-ms moving average filter. The resulting signal was then scaled in order to have a maximum value of one across its length.

The boundaries of the syllables were then computed from the resulting envelope by minima detection: Any local minimum preceded by an amplitude difference larger than δ=0.015 was considered a syllable boundary. In the case where two or more boundaries were closer than 80-ms to each other, a single boundary was considered at the midpoint. Finally, the syllabic nucleus was marked at the local maximum in the envelope between the detected syllable boundaries (see also Räsänen, Doyle, & Frank, submitted, for a comparison with other methods).

Statistical descriptors The main statistical parameters that were used in the analysis of the data were calculated over the duration of each individual word (see, e.g., Chen, Robb, Gilbert, & Lerman, 2001). According to the literature, the most common acoustic descriptors are the mean, median, and variance of the features (see, e.g., Rosenberg, & Hirschberg, 2009; Chen et al., 2001; Eriksson, Barbosa, & Akesson, 2013; Zhang, Nissen, & Francis, 2008). Furthermore, the examination of the maximum and feature change might also provide meaningful information (see, e.g., Terken, 1991). All these measures were calculated over all features in order to gain an understanding of their behavior. Therefore we included the following in the analysis: (i) feature change computed according to Eq. (4), (ii) maximum feature value during the word, (iii) mean during the word, and (iv) variation calculated as the standard deviation of the feature during the word (see also Table 1).

$$f_\psi^{ch} = \max\left\{f_\psi'(t)\right\} - \min\left\{f_\psi'(t)\right\}, t \in [t_1, t_2] \qquad (4)$$

Statistical models

After the computation of the statistical parameters, each word in the vocabulary has a set of descriptors defining the typical behavior of the prosodic features for that word. Each descriptor was then modeled using a normal distribution and a histogram-based probability distribution. While the former provides a first approximation of the typicality of the feature values, the latter can account for any arbitrary-shaped distribution given our present data set with a large number of samples for each word (see experiments).

Table 1: Overview of the statistical parameters used in the experiments.

Features used	Description	Features used	Description
F0_AV	F0 mean	ST_AV	Spectral tilt mean
F0_SD	F0 standard deviation	ST_SD	Spectral tilt standard deviation
F0_CH	F0 change	ST_CH	Spectral tilt change
F0_MX	F0 max	ST_MX	Spectral tilt max
EN_AV	Energy mean	DU_W	Word duration
EN_SD	Energy standard deviation	DU_S	Syllable duration
EN_CH	Energy change		
EN_MX	Energy max		

The normal distribution ($N(\mu,\sigma^2)$) for lexeme L is defined as:

$$\phi_L(f_{\psi,\kappa}, \mu_{\psi,\kappa}, \sigma_{\psi,\kappa}) = \frac{1}{\sigma_{\psi,\kappa}\sqrt{2\pi}} e^{-\frac{(f_{\psi,\kappa}-\mu_{\psi,\kappa})^2}{2(\sigma_{\psi,\kappa})^2}} \qquad (5)$$

where κ denotes the statistical parameter, ψ the acoustic feature, μ the mean value, and σ the standard deviation of the descriptor in the training data. Therefore, for each word L in the vocabulary there are a total of 14 models. It is important to note that, the assumption of normality of the parameters might not necessarily hold for all the examined descriptors, but it is a simple first approach in demonstrating the perceptual effect of deviant features.

During the testing stage, the score for the j:th word token w_{ij} in utterance i for features ψ and descriptors κ was then determined according to Eq. (6):

$$S(w_{ij}) = \sum_{\psi,\kappa} \log_{10}\left[\phi_L(f_{\psi,\kappa,i,j}, \mu_{\psi,\kappa}, \sigma_{\psi,\kappa})\right] \qquad (6)$$

where f denotes the computed feature parameter and L is the known identity of the word. This formulation assumes statistical independence of the feature descriptors in order to combine the individual descriptors and study potential interactions.

The histogram-based distribution model was generated by dividing the descriptor values of the training data into Q uniformly spaced bins across the entire value range. Then each bin was assigned with a probability by taking the proportion of data points that end up in each bin. During testing, the probability of a given feature value was simply the probability of the bin that it was assigned to, while combination of multiple descriptors was performed as a sum of logarithms similarly to the normal distribution in Eq. (6).

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

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

where the threshold was defined at the utterance level as:

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

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

Experiments

The performance of VAP was tested on continuous English speech. To evaluate algorithmic output, a corpus with hand labeled prosodic labels was used (see, Altosaar et al., 2010; Kakouros & Räsänen, 2014b). The annotations were compared against the prominence hypotheses generated by the VAP algorithm. Overall system performance was evaluated using standard measures for accuracy and inter-rater agreement that are further described below.

Material

The CAREGIVER Y2 UK corpus (Altosaar et al., 2010) was used in the experiments reported in this work. The style of speech in CAREGIVER is acted infant-directed speech (IDS) spoken in continuous UK English, simulating a situation where a caregiver is talking to a child 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 Altosaar et al., 2010, for details).

In overall, CAREGIVER Y2 UK contains 2397 sentences from each main talker (approximately 1.8 hours of acoustic data per speaker). Prominence labels were available for a certain part of the corpus (see Kakouros & Räsänen, 2014b, for more information), and therefore, the whole annotated subset of 300 unique utterances was chosen for the purpose of this experiment from one male and female talker (*Speakers 3 and 4*), yielding a total of 600 utterances. The database also contains orthographic transcriptions corresponding to each utterance with time-aligned information at the word level.

All single-word utterances were excluded from the data, leading to an average of 5.9 words per sentence. This set of utterances is referred to as the *test set*, as it was used to probe the performance of the studied VAP model.

Regarding the training of the statistical model, 2000 sentences per talker were used (i.e., 4000 in total) yielding an average of 348 data points per lexeme (SD = 274, median = 212) to estimate the features. None of the utterances in the training set were present in the above test set.

Evaluation

Two standard evaluation approaches were used in order to measure the performance of the VAP: (i) inter-annotator agreement rates and (ii) relevance measures. Specifically, to measure the inter-annotator agreement between the test set and the algorithmic output the standard Fleiss kappa (FK) (Fleiss, 1971) measure was used. FK measures the degree of agreement between two or more annotators on a nominal scale of [-1,1] by taking into account the distribution of the ratings. Therefore, FK yields zero in the case when the distribution is what would be expected if all raters made their judgments completely randomly. In the current work, FK was measured on the word-level. The overall agreement rate on the words in the test set was then used as the primary measure in the analysis. As the prominent labels were available per word from thirteen annotators, a single reference was constructed yielding 1 for prominent and 0 for non-prominent words where a prominence marking was generated for the majority agreement (≥6 votes). FK was then computed between the reference and the prominence hypotheses generated by the algorithm.

For the relevance measures, precision (PRC), recall (RCL), their harmonic mean (F-value), and accuracy (ACC) were used and were defined as:

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

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

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

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

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

Finally, we compare the current approach with the earlier model based on the temporal unpredictability of the feature trajectories. In the feature trajectory model (FTM) (Kakouros & Räsänen, 2014a), the raw acoustic feature of interest (e.g., F0) is first quantized into a finite number of discrete states. An n-gram model is then used to model the typical behavior of these state-sequences over time similarly to language models in speech recognition. Whenever the probability of the prosodic trajectory during an underlying word falls below a pre-defined threshold, it is hypothesized that the word is prominent. In contrast to the present work, the model does not use word information during training, but simply decodes probability information in word-sized temporal chunks during the testing stage (see Kakouros & Räsänen, 2014a, for details).

Experimental setup

The experiment was run by populating the dictionary with all the words and their descriptors available in the training set of 4000 utterances and computing their corresponding statistical models. Next, the proposed VAP approach was tested on 600 novel utterances (test set) where the algorithm selected the words with the most deviant descriptors in each utterance and marked them as prominent. The hyperparameter λ was used in order to control the sensitivity of the algorithm and was set to the range of $\lambda \in [-0.5,1]$.

Results

The experiment was performed for all 14 individual statistical descriptors and also for several combinations and separately for the standard Gaussian (SG) and the histogram (HS) models. Here, only five combinations are considered (see Table 2), as there were many potential feature groupings that would require a more extensive presentation but did not show notable differences from those shown here. The selected subset of the combined descriptors represented two main cases: (i) combination of all feature level descriptors (e.g., all energy descriptors – EN_ALL) and (ii) combination of best performing descriptors – F0_CH, EN_CH, DU_W).

Using the SG, it can be seen (Table 2) that the best performing features were those of duration, energy, and F0 (in decreasing order of performance). Specifically, deviant word durations seemed to be the most descriptive of prominence with $\mathrm{FK}_{\mathrm{SG}}$=0.34 and 76% accuracy (for $\lambda=0$).

Table 2: Results for the individual features and feature combinations for the standard Gaussian (SG – λ=0) and histogram model (HS – Q=3, λ=-0.1).

Features	FK_{SG}	ACC_{SG}	FK_{HS}	ACC_{HS}	Features	FK_{SG}	ACC_{SG}	FK_{HS}	ACC_{HS}
F0_AV	-0.16	0.64	0.12	0.71	ST_CH	-0.17	0.64	0.03	0.69
F0_SD	0.05	0.69	0.15	0.73	ST_MX	-0.27	0.62	-0.24	0.63
F0_CH	0.17	0.71	0.15	0.72	DU_W	0.34	0.76	0.30	0.76
F0_MX	-0.10	0.65	0.12	0.71	DU_S	0.20	0.72	0.10	0.71
EN_AV	-0.08	0.66	0.31	0.76	EN_ALL	0.14	0.71	0.31	0.75
EN_SD	0.08	0.70	0.32	0.75	F0_ALL	0.06	0.69	0.17	0.73
EN_CH	0.21	0.72	0.25	0.74	ST_ALL	-0.29	0.62	-0.05	0.66
EN_MX	0.19	0.72	0.24	0.73	DU_ALL	0.29	0.74	0.25	0.74
ST_AV	-0.22	0.63	0.01	0.67	F0_CH,EN_CH, DU_W	0.31	0.75	0.26	0.74
ST_SD	-0.23	0.63	0.17	0.72					

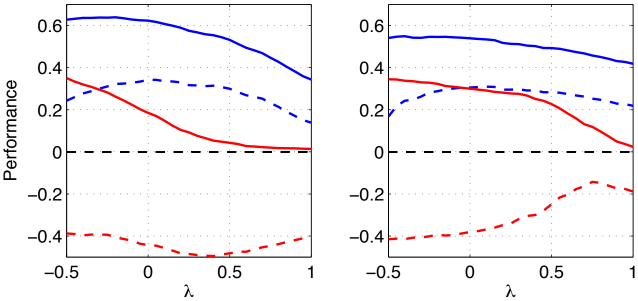

Figure 2: Results for word duration for different values of λ. Left panel: standard Gaussian model. Right panel: histogram-based model. Blue solid line: F-score for deviant DU_W, blue dashed line: FK for deviant DU_W, red solid line: F-score for most expected DU_W, red dashed line: FK for most expected DU_W.

Also, the accuracy level is close to that of a similar study in prominence detection (80%, see Tamburini & Caini, 2005), where, however, a direct comparison is not possible due to the use of different speech corpora. In order to further evaluate VAP, we reversed the conditions in the setup and probed the algorithm to select the words with the most expected descriptors as prominent (highest probability). The results for word duration can be seen in Fig. 2 where it is evident that deviant durations have an effect on prominence perception with FK_{SG}=0.34 (F_{SG}=62%) as opposed to FK_{SG}=-0.48 (F_{SG}=4.6%) for non-deviants.

Next, we run the HS approach for a number of different bin partitions ranging from 2 until 100 in order to find the optimal partition of the probability space. Interestingly, the best partition was obtained for Q=3 and the results are presented in Table 2. In this case, energy seemed to be the best performing feature (FK_{HS}=0.32) followed by word duration (FK_{HS}=0.3) and F0 (FK_{HS}=0.15). Finally, we also run the earlier proposed FTM approach on the same set of data (for 2-grams) that produced FK_{FTM,W_DUR}=0.58 as opposed to FK_{VAP,SG,W_DUR} =0.34 for word duration, $FK_{FTM,F0}$=0.60 as opposed to $FK_{VAP,SG,F0\_CH}$=0.17 for F0 and F0 change respectively and $FK_{FTM,EN}$=0.65 as opposed to FK_{VAP,SG,EN_CH}=0.21 for energy and energy change respectively.

Since the histogram model was overall much better than the Gaussian model (mean FK 0.16 vs. 0.03), we tested for the normality of the descriptor distributions using the Kolmogorov-Smirnov test. The results showed that the majority of the feature descriptors do not follow a normal distribution, confirming that a model consisting of a single Gaussian distribution is simply not suitable for capturing the expectations of prosodic features during words. In contrast, the histogram-based results should be consistent due to the large number of tokens for each word.

In all, the present results are somewhat surprising, suggesting that the lexeme-level predictability of prosodic features does not seem to have a clear function in the perception of sentence prominence as the agreement levels are notably below those obtained using the FTM model.

Discussion and conclusions

The goal of the present study was to investigate whether the predictability of the acoustic correlates of prosody at the level of individual lexemes carries information regarding sentence prominence. Given the earlier finding that sentence level prominence is driven by the unpredictability of prosodic features (see Kakouros & Räsänen, 2014a, for a model on F0), it was of interest whether information regarding the identity of the underlying lexeme would improve from the earlier model by providing more accurate characterization of typical and atypical prosody during the words. The results show that prosodic unpredictability, when conditioned by the lexical content, provides some cues for sentence prominence (see also Aylett & Bull, 1998, for a study using only durational information) but the agreement rates are substantially lower than those produced when using a model that measures (un)predictability of the feature trajectories at the utterance level. Therefore the present results do not show substantial benefits for the role of lexical identity in prominence perception within the predictability framework.

One plausible explanation for this negative finding is that prominence is an utterance level process, and therefore investigating characteristics of individual words in isolation of their sentential context is not meaningful. While the more successful but word-agnostic FTM model (Kakouros & Räsänen, 2014a) analyzes probabilities of prosodic trajectories in a sentential context, the present investigation only looked at word-level aggregate statistics. Moreover, the presently used statistical descriptors were computed across the entire word tokens and this may lose some of the microprosodic information that is not removed in the FTM.

However, it is not presently possible to conclude that lexical information would be irrelevant since there is the possibility that our models simply do not capture the relevant information from the signals. In future work, it would be of interest to find more effective ways of combining lexical knowledge to the predictability framework, possibly augmenting the temporally evolving FTM model instead of directly looking at lexeme-specific statistics. In addition, the present findings should be verified with more sophisticated statistical models such as using Gaussian Mixture Models. Finally, different criteria for aggregating statistics across word tokens could be used, such as part-of-speech (POS) tags, word positions, or syllables rather than the position-invariant lexemes used in the present study.

Acknowledgments

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".

References

Altosaar, T., ten Bosch, L., Aimetti, G., Koniaris, C., Demuynck, K., & van den Heuvel, H. (2010). A Speech Corpus for Modeling Language Acquisition: CAREGIVER. *Proceedings of the International Conference on Language Resources and Evaluation (LREC)* (pp. 1062–1068), Malta.

Aylett, M. P., & Bull, M. (1998). The automatic marking of prominence in spontaneous speech using duration and part of speech information. *Proceedings of International Conference on Spoken Language Processing* (*ICSLP-98*), Sydney, Australia.

Aylett, M., & Turk, A. (2004). The smooth signal redundancy hypothesis: a functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech*, 47(1), 31–56.

Calhoun, S. (2007). Predicting focus through prominence structure. *Proceedings of Interspeech* (pp. 622–625), Antwerp, Belgium.

Calhoun, S. (2010). The Centrality of Metrical Structure in Signaling Information Structure: A Probabilistic Perspective. *Language*, 86(1), pp. 1–42.

Chen, Y., Robb, M. P., Gilbert, H. R., & Lerman, J. W. (2001). A study of sentence stress production in Mandarin speakers of American English. *Journal of the Acoustical Society of America*, 109(4), 1681–1690.

Cole, J., Yoonsook, M., & Hasegawa-Johnson, M. (2010). Signal-based and expectation-based factors in the perception of prosodic prominence. *Laboratory Phonology*, 1(2), 425–452.

Cutler, A., Dahan, D., & van Donselaar, W. (1997). Prosody in the comprehension of spoken language: A literature review. *Language and Speech*, 40, 141–201.

Eriksson, A., Barbosa, P. A., & Akesson, J. (2013). The acoustics of word stress in Swedish: A function of stress level, speaking style and word accent. *Proceedings of Interspeech* (pp. 778–782).

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

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

Itti, L., & Baldi, P. (2009). Bayesian surprise attracts human attention. *Vision Research*, 49, 1295–1306.

Kakouros, S., & Räsänen, O. (2014a). Statistical Unpredictability of F0 Trajectories as a Cue to Sentence Stress. *Proceeding of the 36th Annual Conference of the Cognitive Science Society* (pp. 1246–1251), Quebec, Canada.

Kakouros, S., & Räsänen, O. (2014b). Perception of Sentence Stress in English Infant Directed Speech. *Proceedings of Interspeech* (pp. 1821–1825), Singapore.

Kalinli, O., & Narayanan, S. (2009). Prominence detection using auditory attention cues and task-dependent high level information. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(5), 1009–1024.

Kochanski, G., Grabe, E., Coleman, J., & Rosner B. (2005). Loudness predicts prominence: Fundamental frequency lends little. *Journal of the Acoustical Society of America*, 118(2), 1038–1054.

Lieberman, P. (1960). Some acoustic correlates of word stress in American English. *Journal of the Acoustical Society of America*, 32(4), 451–454.

Mancas M., Beul, D.D., Riche, N., & Siebert, X. (2012). Human Attention Modelization and Data Reduction. In: Intech, ed. Intech., *Video Compression*.

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

Ortega-Llebaria, M., & Prieto, P. (2010). Acoustic correlates of stress in central Catalan and Castilian Spanish. *Language and Speech*, 54(1), 1–25.

Pan, S., & Hirschberg, J. (2000). Modeling local context for pitch accent prediction. *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics* (pp. 233–240).

Rosenberg, A., & Hirschberg J. (2009). Detecting Pitch Accents at the Word, Syllable and Vowel Level. *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL* (pp. 81–84).

Räsänen, O., Doyle, G., & Frank, M. C. (2015). Unsupervised word discovery from speech using automatic segmentation into syllable-like units. *Manuscript submitted for publication*.

Sluijter, A. M. C., & van Heuven, V. J. (1996). Spectral balance as an acoustic correlate of linguistic stress. *Journal of the Acoustical Society of America*, 100(4), 2471–2485.

Tamburini, F., & Caini, C. (2005). An automatic system for detecting prosodic prominence in American English continuous speech. *International Journal of Speech Technology*, 8, 33–44.

Terken, J. (1991). Fundamental frequency and perceived prominence of accented syllables. *Journal of the Acoustical Society of America*, 89(4), 1768–1776.

Tsiakoulis, P., Potamianos, A., & Dimitriadis, D. (2010). Spectral moment features augmented by low order cepstral coefficients for robust ASR. *IEEE Signal Processing Letters*, 17(6), 551–554.

Turk, A. (2010). Does prosodic constituency signal relative predictability? A Smooth Signal Redundancy hypothesis. *Laboratory Phonology*, 1(2), 227–262.

Wagner, P. (2005). Great expectations – introspective vs. perceptual prominence ratings and their acoustic correlates. *Proceedings of Interspeech* (pp. 2381–2384).

Wang, D., & Narayanan, S. (2007). An acoustic measure for word prominence in spontaneous speech. *IEEE Transactions on Audio, Speech, and Language Processing*, *15*(2), 690-701.

Werner, S., & Keller, E. (1994). 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, 23–40.

Zahorian, S. A., & Hu, H. (2008). A spectral/temporal method for robust fundamental frequency tracking. *Journal of the Acoustical Society of America*, 123, 4559–4571.

Zhang, Y., Nissen, S. L., & Francis, A. L. (2008). Acoustic characteristics of English lexical stress produces by native Mandarin speakers. *Journal of the Acoustical Society of America*, 123(6), 4498–4513.