INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Analyzing the Contribution of Top-down Lexical and Bottom-up Acoustic Cues in the Detection of Sentence Prominence

Sofoklis Kakouros[1], *Joris Pelemans*[2], *Lyan Verwimp*[2], *Patrick Wambacq*[2], *Okko Räsänen*[1]

[1] Department of Signal Processing and Acoustics, Aalto University, Finland
[2] ESAT, KU Leuven, Belgium

{sofoklis.kakouros,okko.rasanen}@aalto.fi,
{joris.pelemans,lyan.verwimp,patrick.wambacq}@esat.kuleuven.be

Abstract

Recent work has suggested that prominence perception could be driven by the predictability of the acoustic prosodic features of speech. On the other hand, lexical predictability and part of speech information are also known to correlate with prominence. In this paper, we investigate how the bottom-up acoustic and top-down lexical cues contribute to sentence prominence by using both types of features in unsupervised and supervised systems for automatic prominence detection. The study is conducted using a corpus of Dutch continuous speech with manually annotated prominence labels. Our results show that unpredictability of speech patterns is a consistent and important cue for prominence at both the lexical and acoustic levels, and also that lexical predictability and part-of-speech information can be used as efficient features in supervised prominence classifiers.

Index Terms: sentence prominence, prosody, stimulus predictability, speech perception

1. Introduction

Prosody plays a central role in spoken communication, corresponding to the manner in which words are spoken and comprising information that may not be available in the lexical content of sentences. Prosodic prominence is a particularly important prosodic phenomenon that can be described as the property by which a linguistic entity is perceived to be standing out from its environment [1]. More specifically, sentence prominence can be generally defined as describing the perceived emphasis of one or more words during a sentence (see also [2–4] for related definitions).

The role of prominence in discourse can be identified at many levels, making its production and interpretation critical for spoken communication. For instance, prominence can be indicative of information structure or lexical class (see [5] and references therein). Thus, methods for automatic prominence detection can have various uses in spoken language applications, such as during the development of text-to-speech (TTS) systems where it is particularly important to achieve a naturalistic production of speech (see, e.g., [6,7]). Similarly, there are various applications based on automatic speech recognition (ASR) systems such as that of spoken content retrieval [8] and topic tracking [9].

It is well known that sentence prominence is highly correlated with prosodic acoustic features such as energy, fundamental frequency, and word duration. Recent work by Kakouros and Räsänen [10] suggests that perception of prominence might result from surprisal (unpredictability) in the suprasegmental features of speech instead of depending on certain ("fixed") feature patterns or magnitudes as such. However, earlier work also indicates that predictability at the lexical level is related to word prominence (see, e.g., [11]), suggesting that prominence is a phenomenon that is reflected at multiple levels of speech and language.

In this paper, we analyze both bottom-up acoustic cues and top-down lexical cues in the task of automatic prominence detection. More specifically, we study both types of cues in isolation, and in combination, using both unsupervised and supervised systems for prominence detection. As a result, we show how predictability at multiple levels contributes to the impression of prominence in speech.

1.1. Background

Prosody is a property of speech that can be viewed from both the linguistic as well as the phonetic (acoustic) perspective. At the linguistic level, prosody refers to the phonological organization of segments into higher-level constituents (for instance, syllables, words, sentences) and the prosodic phenomena superimposed on them (for instance, prominence) (see, e.g., [12], for related definitions). At the phonetic level, prosody refers to the acoustic parameters that can best describe the prosodic events observed at the linguistic level (see, e.g., [12]). The latter also defines the physical correlates associated with the acoustic realization of prominence. Therefore, the variations in the acoustic domain that manifest perceptual differences with respect to prominence are typically the fundamental frequency (F0) (e.g., [13]), duration (see, e.g., [14]), energy [15], and spectral tilt [16] (see also [17]). In the case of sentence prominence, these variations result in relative differences in the perceived salience between the words in a sentence. As saliency is closely connected to perceptual and attentional orientation (see, e.g., [18] and references therein), an analogy can be drawn between prominence perception and the cognitive function of attention [19].

Saliency and attention can be computationally modeled based on a probabilistic formulation where low-probability, i.e., surprising, events are considered as prominent (see, e.g., [20]). In the context of speech, this means that we can model prominence by learning a model for typical prosodic features or feature trajectories from a set of unlabeled speech data and then evaluate the overall predictability of these features over time (see, e.g., [21,22]). Frequency and predictability effects are also known to play an important role in models of human language production and comprehension [23], thus, utilizing predictability at different levels of analysis in speech can be also very important in modeling prominence (see, e.g., [10,11]).

http://dx.doi.org/10.21437/Interspeech.2016-926

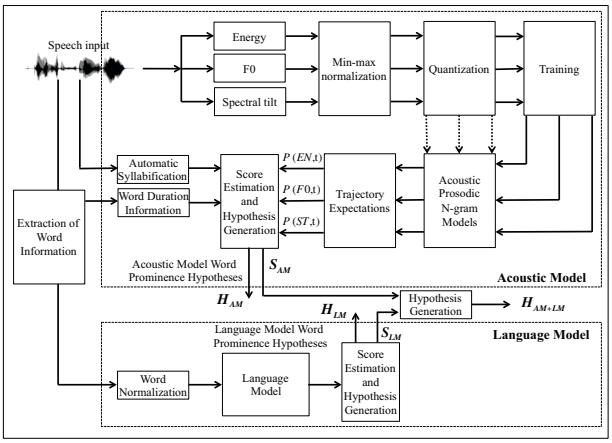

Figure 1: *Schematic diagram of the processing steps of the unsupervised algorithm consisting of an acoustic model and a language model and where H denotes the word hypothesis and S the word score.*

In earlier studies, predictability at the level of individual lexemes has been shown to be a consistent indicator of pitch accent placement in speech [11]. Specifically, Pan and McKeown [11] evaluated the relative informativeness of words using measures such as the negative log-likelihood of a word in a corpus and found that there is a positive correlation between the informativeness of a word and its pitch accent placement. In a later study, Pan and Hirschberg [24] assessed the effect of word collocation information and accent placement using measures such as the log-conditional predictability and mutual information. Their results indicate that the more predictable a word is based on its local context, the more likely it is to be deaccented (see also [25]).

In all, it seems that word predictability can be used for the prediction of prominence, but the interactions between lexical predictability and acoustic predictability are currently unclear. So far, according to our knowledge, the only computational algorithms combining both lexical (predictability) and acoustic cues are the studies of Rosenberg et al. [26] and Fernandez and Ramabhadran ([27]) that used bigram word probabilities and standard acoustic features for supervised prominence detection. However, those studies did not investigate the relative impact of different features on the overall performance of the systems.

In the present study, we investigate the contribution of bottom-up acoustic and top-down lexical cues in both supervised and unsupervised systems for prominence detection. Supervised methods provide a performance benchmark for the different features being evaluated and are expected to lead to the best overall performance. The unsupervised method provides a cognitively plausible proposal based on the predictability framework described in [22] that can be used for prominence detection without prosodically labeled training data. The results are evaluated on a set of annotated data of Dutch continuous speech with performance of the unsupervised method being close to that of supervised methods on the same task.

2. Methods

A number of acoustic and lexical features were computed and used in an unsupervised (section 2.2) and supervised setups (section 2.3) for prominence detection. In the unsupervised approach, no prominence labels are used in the training of the system. Instead, the algorithm evaluates the overall unpredictability of speech input at both the acoustic and lexical levels on the basis of two models: (1) a language model (LM) that provides the predictability of each word in a certain lexical context and (2) an acoustic model (AM) for prosodic features, providing frame-by-frame estimates of prosodic predictability (Figure 1). As a result, words with low acoustic and/or lexical predictability are considered as more prominent than highly predictable words. In the supervised setup, two standard classifiers, support vector machines (SVMs) and k-nearest neighbors classifier (KNNs) were tested using word-level acoustic descriptors and language model probabilities as well as part-of-speech tags as features in the classification.

2.1. Features

2.1.1. Acoustic features

F0, energy (EN), spectral tilt, and duration (dur) were used as the main acoustic features. For this purpose, speech data were first downsampled to 8 kHz. F0 estimation was carried out using the YAAPT algorithm [28] with a 25-ms window and 10-ms step size. The pitch tracks were linearly interpolated across unvoiced segments in order to preserve continuity of the features. Energy was computed similarly in 25-ms windows using a 10-ms step size, and spectral tilt was extracted as the first Mel-frequency cepstral coefficient from standard MFCC computation using the same windowing procedure [29]. Word durations were extracted from the annotations of the Dutch corpus while syllable durations were automatically estimated using the sonority-envelope based algorithm described in [30].

For supervised classification purposes, five word-level statistical descriptors were calculated for F0, EN, and tilt, namely: (i) mean, (ii) max, (iii) min, (iv) variance, and (v) the mean first-order difference during the word. These word-level feature descriptors were only used in the supervised classification scenario while the original continuous-time F0, EN, and tilt were used in the unsupervised model (section 2.2).

2.1.2. Lexical features

Part-of-speech tags (POS) and word *n*-gram statistics were used as the two primary lexical features. POS tags were extracted directly from the corpus annotations while *n*-gram statistics from unigrams to 5-grams were computed using the Dutch language models described in the next sub-section (2.2.1). Log-probabilities for all five *n*-gram orders were used as features for each word, thereby quantifying the relative surprisal of each word given the preceding context of *n-1* words. For supervised classification, the POS tags (13 unique classes) were expanded into 13-dimensional binary feature vectors for each word, one dimension for each POS class.

2.2. Unsupervised prominence detection system

2.2.1. Language (top-down) model

Five Dutch LMs were used in our model to provide word predictability estimates. The training material for the LMs consisted of newspaper and magazine articles (1.3B word tokens in total), that were pre-processed in three steps for the

purpose of speech recognition: normalization, spelling correction and filtering (see [31]). During text normalization, punctuation was used to determine sentence splits and was then removed; words containing a number and an alphabetic segment (e.g. *100-jarige*) were split; numbers, measure words, web addresses, abbreviations, etc., were written in full and text segments that were likely to be trash were removed. In the next step, incorrectly capitalized words (at the beginning of a sentence) were corrected if the word occurred more often with lowercase than with uppercase in the corpus, and words with a hyphen were split (based on inter-corpus statistics). Spelling variants were detected by looking for words that had comparable bigram statistics and a maximum Levenshtein distance of two. These were also manually verified. In the filtering step, duplicate sentences along with sentences with many uncommon words were removed.

The pre-processed data were then used to train *n*-gram LMs using the SRILM toolkit [32]. We trained models with *n*-gram orders ranging from 1 to 5, a vocabulary of the 400k most frequent words and count cut-offs of 1 (for 1- and 2-grams) and 2 (for 3-, 4- and 5-grams). All models were smoothed using modified Kneser-Ney [33], except for the unigram model that was smoothed with Good-Turing [34].

2.2.2. Acoustic (bottom-up) model

Analogously to the language model, the acoustic model provides the probabilities of the acoustic prosodic trajectories and is based on the work described in [10]. Specifically, for energy, F0, and spectral tilt, the raw features were initially min-max normalized and quantized to $Q = 16$ discrete levels using the k-means algorithm. The discretized feature values were then used to train *n*-gram models of different orders. The probabilities $P'(t)$ of the discrete n-tuples at time t (a_t,… a_{t-n+1}) were then computed by summing the log-probabilities over the features (ψ) of interest (Eq. (1)). The resulting frame probabilities were modulated based on syllable duration $d(t)$ and acoustic word scores S_{AM} were computed according to Eq. (2) where t_{start} and t_{end} denote the start and end of the word, as extracted from the annotation (see [10] for a more detailed description).

$$P'(t) = \sum_{\psi} \log_{10}\left(P_{\psi}(a_t \mid a_{t-1}, ..., a_{t-n+1})\right) \quad (1)$$

$$S_{AM}(w_{ij}) = \sum_{t=t_{\mathrm{start}}}^{t_{\mathrm{end}}} P'(t) \times e^{d(t)} \quad (2)$$

Note that syllable durations $d(t)$ are measured in seconds, and therefore $e^{d(t)}$ results in a nearly linear scaling for typical syllable lengths with a min value of 1 for very short syllables.

2.2.3. Prominence hypotheses generation

The prominence hypothesis $H(w)$ for each word i in utterance j was evaluated based on whether the word-level score ($S(w_{ij})$) falls below a detection threshold r_i:

$$H(w_{ij}) = \begin{cases} 1, & S(w_{ij}) < r_i, \\ 0, & S(w_{ij}) \geq r_i \end{cases} \quad (3)$$

$$r_i = \mu_i - \sigma_i \lambda \quad (4)$$

The threshold is defined locally at the utterance level based on the mean (μ) and standard deviation (σ) of the word-level scores across the utterance. The hyperparameter λ controls the overall sensitivity of the detection process. For the top-down lexical model, the word score was represented by $S_{LM}(w_{ij}) = \log_{10}(\mathrm{P}(w_{ij} \mid w_{i-1,j},\ldots,w_{i-n+1,j})$, i.e., the language model output for *n*-gram order *n*, while for the bottom-up case, the corresponding score is $S_{AM}(w_{ij})$ in (Eq. (6)). Finally, for their combination, the summed scores represent the contributions of both models, that is, $S_{AM+LM}(w_{ij}) = S_{AM}(w_{ij}) + S_{LM}(w_{ij})$ (note that the scores are in log-probability domain, corresponding to multiplication of probabilities). In the experiments, the resulting detection accuracies are reported for the optimal λ value in each condition.

2.3. Supervised classification

KNN and SVM classifiers were used to analyze the contribution of acoustic and lexical cues in a supervised classification scenario. For KNN, the number of nearest neighbors was set to $k = 13$ since this provided the most consistent performance in preliminary testing. SVMs used radial basis function with a scale factor of $\sigma = 12.08$ and box constraint $C = 100$, as these were previously optimized in the context of acoustic features with the same data [35].

Using the five word level descriptors described in section 2.1.1, all possible combinations of energy, F0, spectral tilt, word probability, and POS features were tested separately with both classifiers. All five *n*-gram orders were always used as a feature vector for lexical predictability of each word.

3. Experiments

3.1. Material

The Spoken Dutch Corpus (Corpus Gesproken Nederlands; CGN) was used in our experiments. CGN is a database of contemporary standard Dutch as spoken by adults in the Netherlands and Flanders containing nearly 9 millions words (800 hours of speech). The database includes manually generated or verified annotations such as phonetic transcriptions, word level alignment, and prosodic annotations (see [36] for a more detailed description). The Dutch news broadcast ("*component k*") section of the corpus was used in the current experiments, consisting of 5088 news broadcasts (≈27.4 hours of speech data) spoken by 29 speakers (22 male and 7 female). The prosodically annotated subset of the section consists of 134 news broadcasts spoken by 10 different speakers (9 male and 1 female) (≈44.3 minutes of speech data) that were hand-labeled using binary (prominent/non-prominent) markings by two trained annotators, containing a total of 7438 word tokens.

A 10-fold evaluation process was used in the experiments. In the unsupervised case, the full broadcast section plus nine talkers from the prosodically annotated section were always used for training while the remaining talker from the annotated subset was used for evaluation. For the supervised condition, only the labeled section of the corpus was used for training.

3.2. Evaluation

All evaluations were carried out at the word level, i.e., comparing the manually labeled prominence markings to the word-level hypotheses provided by the algorithms, both represented as binary decisions for the presence or absence of prominence in the words. Precision (PRC), recall (RCL), their harmonic mean (F-value), and accuracy (ACC) were used as the main measures:

$$\mathrm{RCL} = tp / (tp + fn) \quad (5)$$

$$\mathrm{PRC} = tp / (tp + fp) \quad (6)$$

$$\mathrm{F} = (2 \times \mathrm{PRC} \times \mathrm{RCL})/(\mathrm{PRC}+\mathrm{RCL}) \quad (7)$$

$$\mathrm{ACC} = (tp + tn) / (tp + fp + fn + tn) \quad (8)$$

where *tp* denotes the true positives, *tn* the true negatives, *fp* the false positives, and *fn* the false negatives. In addition, we report results using Fleiss kappa that measures the degree of agreement between two or more annotators on a nominal scale of $\kappa \in [-1, 1]$. Fleiss kappa yields $\kappa = 0$ if the agreement is equal to chance-level co-occurrences in the data and $\kappa = 1$ if all annotators fully agree. Finally, we also report accuracy for the supervised system (% of tokens correctly classified) since this is commonly used in the literature, although accuracy is not very informative in comparison of results between different corpora due to its sensitivity to the underlying class distribution in the data. All results are computed using an annotation reference including all words that either or both of the annotators labeled as prominent.

3.3. Results

For the unsupervised system and for each fold, five orders of the acoustic *n*-gram models ($n = 1, 2, 3, 4$, and 5) were trained for energy, F0, and spectral tilt, on speech data from 28 speakers, always keeping 1 (out of 10) of the annotated speakers for evaluation. In order to evaluate performance for different threshold levels, hyperparameter λ was varied between [-2, 2] with steps of 0.05 for the lexical, acoustic, and combined models (see [10,35] for examples on the effect of λ). Table 1 presents the results for the independent features and the most relevant combinations, as well as the combined model (acoustic+lexical) performance. For the acoustic model, results are shown only for the acoustic bigrams as they were the best performing *n*-gram order.

Both the acoustic and lexical models independently reached high performance in prominence detection in the unsupervised system. The acoustic model alone reached ACC = 86% based on a combination of EN and F0, while the lexical model also performed well, achieving ACC = 82% using unigrams. Higher orders of the lexical *n*-grams seem to deteriorate the performance of the lexical model. Overall, it seems that predictability of the speech stream is a strong cue for prominence at both the lexical and acoustic levels. However, the combination of lexical and acoustic information adds only a small increase in the overall system performance, leading to an accuracy of 87%. Note that the performance of the unsupervised system is significantly higher than that of other unsupervised methods reported in the literature (ACC = 78.1% in [18] and 80.61% in [37]) using different feature combinations. However, direct comparison is not possible due to the different corpora used in different studies.

Table 2 shows the performance measures for the supervised classifiers in the same task. Only the most relevant feature combinations are shown for the sake of consciseness. Performance of all other combinations was within the performance range of the shown results. As can be observed, both acoustic and lexical features are informative also in the supervised case, acoustic features being slightly better when used alone. F0 seems to be the most informative as a sole feature, closely followed by duration. This is in contrast with the unsupervised system where predictability of energy was in fact more useful than F0, equaling supervised F0-based classification in performance. As for the classifiers, SVM always outperforms or equals to KNN in every condition. In practice, *n*-gram probabilities and word durations had a very strong inverse correlation ($r = -0.81$, rank-cor) and performed similarly in combination with other features. This replicates the well-known finding that common (highly predictable) words also tend to be shorter (e.g., function words).

Table 1: *Results for the individual features and selected best performing combinations in the unsupervised experiments.*

	FEATURES	κ	F	ACC		FEATURES	κ	F	ACC
LEXICAL	1-gram	**0.63**	**0.80**	**0.82**	COMBINED	EN+F0+1-gram	**0.72**	**0.84**	**0.87**
	2-gram	0.51	0.72	0.76		EN+F0+2-gram	0.70	0.83	0.85
	3-gram	0.43	0.68	0.72		EN+F0+3-gram	0.69	0.83	0.85
	4-gram	0.42	0.67	0.71		EN+F0+4-gram	0.69	0.82	0.85
	5-gram	0.42	0.67	0.71		EN+F0+5-gram	0.68	0.83	0.84
	combined	0.50	0.73	0.75		EN+F0+comb.	0.70	0.83	0.85
ACOUSTIC	EN+F0+tilt	0.70	0.83	0.85					
	EN+F0	**0.72**	**0.84**	**0.86**					
	F0	0.67	0.79	0.82					
	EN	0.69	0.83	0.85					
	tilt	0.60	0.82	0.84					

Table 2: *Results for the individual features and selected best performing combinations from the supervised classification experiments.*

		SVM			KNN		
	FEATURES	κ	F	ACC	κ	F	ACC
COMBINED	EN+F0+dur+POS+n-grams	**0.77**	**0.87**	**0.88**	**0.74**	**0.86**	**0.87**
	F0+tilt+POS+n-grams	**0.77**	**0.87**	**0.88**	0.73	0.85	**0.87**
	EN+F0+dur+POS	0.76	**0.87**	**0.88**	**0.74**	0.85	**0.87**
	EN+F0+POS+n-grams	0.76	**0.87**	**0.88**	**0.74**	0.85	**0.87**
LEXICAL	POS+n-grams	0.71	0.84	0.85	0.69	0.83	0.84
	n-grams	0.63	0.80	0.81	0.61	0.79	0.81
	POS	0.62	0.78	0.81	0.62	0.78	0.82
ACOUSTIC	EN+F0+tilt+dur	0.74	0.86	0.87	0.71	0.84	0.86
	EN+F0+tilt	0.73	0.85	0.87	0.67	0.82	0.84
	EN+F0	0.72	0.84	0.86	0.69	0.82	0.85
	F0	0.69	0.82	0.85	0.69	0.82	0.85
	dur	0.67	0.82	0.84	0.66	0.81	0.83
	EN	0.63	0.79	0.82	0.59	0.77	0.80
	tilt	0.60	0.78	0.81	0.57	0.76	0.79

Overall, the supervised performance of the best feature combination is close to the result of ACC = 89.03% reported in the literature using Bidirectional Recurrent Neural networks [26], although, as mentioned, experiments with different corpora are not directly comparable. The performance is also higher than many other recent approaches using a variety of features and classifiers [18,38,39,40,41].

4. Conclusions

This study presented an investigation of the contribution of bottom-up acoustic and top-down lexical features in both unsupervised and supervised setups. Overall, the findings suggest that predictability is a strong cue for prominence at both the lexical and acoustic levels, leading to prominence detection accuracy of 87% at the word level using a system trained without prominence labels. Moreover, a combination of word predictability and part-of-speech information proved to be useful for standard supervised classification of prominence, especially when combined with acoustic features, providing high agreement with manual annotations of prominence in the CGN data. Future efforts will investigate the contribution of bottom-up and top-down cues in different languages.

5. Acknowledgements

This study was funded by the Academy of Finland in the project "Computational modeling of language acquisition", by IWT (project 130041, SCATE), IWT–INNOVATIEF AANBESTEDEN and VRT in the STON project.

6. References

[1] J. Terken and D. Hermes, "The perception of prosodic prominence," in *Prosody: Theory and experiment. Studies presented to Gösta Bruce,* M. Horne, Ed. Dordrecht, The Netherlands: Kluwer, pp. 89–127, 2000.

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

[3] A. Cutler, "Lexical Stress," in *The handbook of speech perception*, D. B. Pisoni and R. E. Remez, Eds. Blackwell publishing, pp. 264–289, 2005.

[4] A. Cutler, D. Oahan, and W. Donselaar, "Prosody in the comprehension of spoken language: A literature review," *Language and Speech*, vol. 40, pp. 141–201, 1997.

[5] P. Wagner et al., "Different parts of the same elephant: a roadmap to disentangle and connect different perspectives on prosodic prominence," in *Proceedings of ICPhS*, 2015.

[6] J. Hirschberg and O. Rambow, "Learning prosodic features using a tree representation," in *Proceedings of INTERSPEECH,* pp. 1175–1178, 2001.

[7] M. Mehrabani, T. Mishra, and A. Conkie, "Unsupervised prominence prediction for speech synthesis," in *Proceedings of INTERSPEECH,* pp. 1559–1563, 2013.

[8] D. N. Racca and G. J. Jones, "Incorporating Prosodic Prominence Evidence into Term Weights for Spoken Content Retrieval," in *Proceedings of INTERSPEECH,* pp. 1378–1382, 2015.

[9] C. Guinaudeau and J. Hirschberg, "Accounting for prosodic information to improve ASR-based topic tracking for TV broadcast news," in *Proceedings of INTERSPEECH*, pp. 1401–1404, 2011.

[10] S. Kakouros and O. Räsänen, "Automatic Detection of Sentence Prominence in Speech Using Predictability of Word-Level Acoustic Features," in *Proceedings of INTERSPEECH,* pp. 568–572, 2015.

[11] S. Pan and K. McKeown, "Word informativeness and automatic pitch accent modeling," in *Proceedings of EMNLP/VLC*, pp. 148–157, 1999.

[12] S. Shattuck-Hufnagel and A. E. Turk, "A prosody tutorial for investigators of auditory sentence processing," *Journal of psycholinguistic research*, vol. 25, no. 2, pp. 193–247, 1996.

[13] J. Terken, "Fundamental frequency and perceived prominence of accented syllables," *Journal of the Acoustical Society of America*, vol. 89, no. 4, pp. 1768–1776, 1991.

[14] P. Lieberman, "Some acoustic correlates of word stress in American English," *Journal of the Acoustical Society of America*, vol. 32, no. 4, pp. 451–454, 1960.

[15] G. Kochanski, E. Grabe, J. Coleman, and B. Rosner, "Loudness predicts prominence: Fundamental frequency lends little," *Journal of the Acoustical Society of America*, vol. 118, no. 2, pp. 1038–1054, 2005.

[16] A. M. C. Sluijter and V. J. van Heuven, "Spectral balance as an acoustic correlate of linguistic stress," *Journal of the Acoustical Society of America*, vol. 100, no. 4, pp. 2471–2485, 1996.

[17] N. Campbell, "Loudness, spectral tilt, and perceived prominence in dialogues," in *Proceedings of ICPhS*, pp. 676–679, 1995.

[18] O. Kalinli and S. 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.

[19] J. Cole, Y. Mo, and M. Hasegawa-Johnson, "Signal-based and expectation-based factors in the perception of prosodic prominence," *Laboratory Phonology*, vol. 1, no. 2, pp. 425–452, 2010.

[20] L. Itti and P. Baldi, "Bayesian surprise attracts human attention," *Vision research*, vol. 49, no. 10, pp. 1295–1306, 2009.

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

[22] S. Kakouros and O. Räsänen, "Perception of sentence stress in speech correlates with the temporal unpredictability of prosodic features," *Cognitive Science* (in press).

[23] D. Jurafsky, "A probabilistic model of lexical and syntactic access and disambiguation," *Cognitive Science*, vol. 20, pp. 137–194, 1996.

[24] S. Pan and J. Hirschberg, "Modeling local context for pitch accent prediction," in *Proceedings of the Annual Meeting of the ACL*, pp. 233–240, 2000.

[25] A. Nenkova, J. Brenier, A. Kothari, S. Calhoun, L. Whitton, D. Beaver, and D. Jurafsky, "To memorize or to predict: Prominence labeling in conversational speech," in *Proceedings of NAACL-HLT*, pp. 9–16, 2007.

[26] A. Rosenberg, R. Fernandez, and B. Ramabhadran, "Modeling Phrasing and Prominence Using Deep Recurrent Learning," in *Proceedings of INTERSPEECH*, pp. 3066–3070, 2015.

[27] R. Fernandez and B. Ramabhadran, "Discriminative training and unsupervised adaptation for labeling prosodic events with limited training data," in *Proceedings of INTERSPEECH,* pp. 1429–1432, 2010.

[28] S. A. Zahorian and H. Hu, "A spectral/temporal method for robust fundamental frequency tracking," *Journal of the Acoustical Society of America*, vol. 123, pp. 4559–4571, 2008.

[29] P. Tsiakoulis, A. Potamianos, and D. Dimitriadis, "Spectral moment features augmented by low order cepstral coefficients for robust ASR," *IEEE Signal Processing Letters*, vol. 17, no. 6, pp. 551–554, 2010.

[30] O. Räsänen, G. Doyle, and M. C. Frank, "Unsupervised word discovery from speech using automatic segmentation into syllable-like units," in *Proceedings of INTERSPEECH*, pp. 3204–3208, 2015.

[31] K. Demuynck, A. Puurula, D. Van Compernolle, and P. Wambacq, "The ESAT 2008 system for N-Best Dutch speech recognition benchmark," in *Proceedings of ASRU*, pp. 339–343, 2009.

[32] A. Stolcke, "SRILM – an extensible language modeling toolkit," in *Proceedings of INTERSPEECH*, pp. 901–904, 2002.

[33] S. F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," *Computer Speech and Language*, vol. 13, pp. 359–394, 1999.

[34] S. M. Katz, "Estimation of probabilities from sparse data for the language model component of a speech recognizer," *IEEE Transactions on Acoustics, Speech and Signal Processing,* vol. 35, no. 3, pp. 400–401, 1987.

[35] S. Kakouros and O. Räsänen, "3PRO – An Unsupervised Method for the Automatic Detection of Sentence Prominence in Speech," *Speech Communication*, vol. 82, pp. 67–84, 2016.

[36] J. Duchateau, T. Ceyssens, and H. Van Hamme, "Use and evaluation of prosodic annotations in Dutch," in *Proceedings of LREC,* pp. 1517–1520, 2004.

[37] 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.

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

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

[40] T. Mishra, V. K. R. Sridhar, and A. Conkie, "Word Prominence Detection using Robust yet Simple Prosodic Features," in *Proceedings of INTERSPEECH*, pp. 1864–1867, 2012.

[41] G. Christodoulides and M. Avanzi, "An evaluation of machine learning methods for prominence detection in French," in *Proceedings of INTERSPEECH*, pp. 116–119, 2014.