Statistical Learning of Prosodic Patterns and Reversal of Perceptual Cues for 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

Recent work has proposed that prominence perception in speech could be driven by predictability of prosodic patterns, connecting prominence perception to the concept of statistical learning. In the present study, we tested the predictability hypothesis by conducting a listening test where subjects were first exposed to a 5-minute stream of sentences with a certain proportion of sentence-final words having either a falling or rising pitch trajectory. After the exposure stage, subjects were asked to grade prominence in a set of novel sentences with similar pitch patterns. The results show that the subjects were significantly more likely to perceive words with low-probability pitch trajectories as prominent independently of the direction of the pitch change. This suggests that even short exposure to prosodic patterns with a certain statistical structure can induce changes in prominence perception, supporting the connection between prominence perception and attentional orientation towards low-probability events in an otherwise predictable context.

Keywords: statistical learning; prosody; prominence perception; attention; stimulus predictability

Introduction

Recent theoretical and computational studies have suggested that there is a connection between perception of prominence and the predictability of the acoustic prosodic features in speech (Kakouros & Räsänen, 2014; 2015a; in press). The general idea is that subjective perception of prominence is a response elicited to unpredictable prosodic trajectories in a normal train of speech, thereby drawing the attention of the listener. This hypothesis extends the existing probabilistic accounts of prominence that are based, for instance, on the frequency of occurrence of linguistic units such as syllables and words (see, e.g., Aylett and Turk, 2004; 2006) or on word collocation information (see, e.g., Pan & Hirschberg, 2000). The proposal presented in Kakouros & Räsänen (in press) attempts to explain how the speaker must be also capable of manipulating the expectations of the prosodic correlates of prominence in the acoustic signal, therefore going beyond the probabilistic relations at the symbolic linguistic level.

In the present study, we conducted a listening experiment where probabilities of prosodic trajectories were explicitly manipulated, first exposing subjects to different ratios of rising and falling pitch trajectories on sentence-final words and then asking the subjects to grade prominence in a set of novel utterances. The experiments and results, as described in the next sections, show that the probability of prosodic features indeed affects subjective perception of prominence.

Background

Prominence is a prosodic phenomenon that can be generally defined as the property by which linguistic units are perceived to be standing out from their environment (Terken, 1991) and is closely connected to the concept of stress. As the terminology can be ambiguous, here we use the term sentence prominence to refer to one or more words that are perceived to be standing out in a sentence (see, e.g., Terken, 1991; Cutler, Dahan, & van Donselaar, 1997, for related definitions).

Prominence has been examined from a number of different perspectives (see, Wagner et al., 2015, and references therein). From the physical perspective, a number of studies have focused on identifying the acoustic correlates of prominence. It has been now well established that energy, fundamental frequency (F0), duration (see, e.g., Fry, 1955; 1958; Lieberman, 1960; Terken, 1991; Kochanski, Grabe, Coleman, & Rosner, 2005; see also Ortega-Llebaria & Prieto, 2010, and references therein), and spectral tilt (see, e.g., Sluijter & van Heuven, 1996; but see also Campbell, 1995; Campbell & Beckman, 1997) are the acoustic parameters in speech whose variations signal prominence and constituent boundaries (see also Shattuck-Hufnagel & Turk, 1996). Another interesting aspect of these features is the degree of acoustic similarities in prominence production and perception across different languages. The central argument is that regardless of the language, all speakers are equipped with the same production and perception apparatus, therefore, the type of information conveyed through speech should not vary greatly (Vaissière, 1983). This assumption may at least be partly true for prominence as it seems that the basic acoustic correlates of prosody are the same, although the actual realizations of prosodic patterns depend on the language (see, e.g., Rosenberg et al., 2012; Maier et al., 2009). Additionally, a speaker can manipulate the acoustic prosodic features relatively independently of the linguistic content of a sentence. On the listener's side, the perceptual outcome of prominence seems to be the same across languages, that is, a

shift of the attention to a specific part in the stream of speech.

From the functional perspective, prominence has been studied with respect to its linguistic and communicative role. The realization of prominence seems to have effects on the parsing of information and syntactic structure of utterances (see, e.g., Calhoun, 2010; Shattuck-Hufnagel & Turk, 1996). For instance, prominence may indicate the word in an utterance where the most important information lies and it has been observed that reaction times (RTs) for prominent words are shorter when compared to their non-prominent counterparts (see, e.g., Cutler & Foss, 1977). In all, prominent words seem to attract the listener's attention thus allocating extra cognitive processing resources (see, e.g., Cole, Mo, & Hasegawa-Johnson, 2010). At the level of the listener's perceptual processing this implies that there are elements in speech that attract the listener' attention. Cole et al. (2010) have suggested that prominence and attention might be associated, where a listener's attention can be drawn to a word either as a response to acoustic modulation or due its relative unpredictability. In general, an attention-capturing stimulus can be seen as something that is novel or surprising (see, e.g., Itti & Baldi, 2009). Correspondingly, surprisal can be defined in a probabilistic way as something that is unpredictable. In the case of the acoustic prosodic features, this could be manifested, for instance, as an unpredictable F0 trajectory (see, e.g., Kakouros & Räsänen, 2014).

Probabilistic processing at the level of human cognition is an idea that has held a central role in many models of language processing. Predictability and frequency effects have been widely studied, providing evidence that predictability plays a role in language comprehension, production, and learning (see, e.g., Jurafsky, Bell, Gregory, & Raymond, 2001; Jurafsky, 1996). For instance, predictability of the linguistic elements (such as syllables or words) seems to affect their acoustic realization during speech production (see, e.g., Jurafsky et al., 2001). For example, frequent words are more likely to be reduced in duration than less frequent words. Several theories have emerged in an attempt to explain these probabilistic phenomena, resulting in theories such as the Probabilistic Reduction Hypothesis (Jurafsky et al., 2001) and Uniform Information Density (Frank & Jaeger, 2008). At the level of prosodic prominence, Aylett and Turk's (2004) Smooth Signal Redundancy Hypothesis is based on a similar proposal (linguistic predictability) and suggests that acoustic differences are linguistically implemented through prosodic prominence structure. However, most of the theories focus on examining the predictability of the linguistic units (e.g., phonemes, words) in speech whereas little is known about how the predictability of the low-level or suprasegmental acoustic features affects speech production and perception. Moreover, not all acoustic variation can be explained only by differences in the predictability at lexical or grammatical level. Thus, investigating how the predictability of the acoustic prosodic features might affect different phenomena in speech production and perception is of particular interest.

Previous computational modeling studies reveal that predictability of prosodic trajectories, when measured in terms of a probabilistic prosody model learned from a corpus of speech, is highly correlated with human perception of prominence in the same set of utterances (Kakouros & Räsänen, 2014; 2015b; in press). In the present work, we investigate whether it is possible to induce different prominence perception patterns in human subjects by manipulating the probabilities of different prosodic trajectories during a habituation stage. More specifically, we ask whether the predictability of F0 trajectories affects the perception of sentence prominence. The hypothesis is that low-probability F0 patterns, i.e., the patterns that are less frequent during the habituation stage, would be regarded as more prominent independently of the actual direction of the F0 change.

Experimental setup

We conducted a listening experiment to investigate whether exposure to a certain probability distribution of rising and falling F0 trajectories will affect the listeners' perception of prominence. The aim was to examine 1) whether probabilities of prosodic trajectories have an impact on subjective prominence ratings, and 2) whether a short exposure to prosodic stimuli is sufficient to alter these probabilistic expectations from the baseline perceptual system acquired through life-long experience with spoken language.

The experiment consisted of two conditions: The first, referred to as rising standard condition (RSC), involved the presentation of spoken utterances with 90% of the tokens having rising fundamental frequency (F0) during the sentence-final word while 10% of the sentences had falling F0 during the last word. The second condition, the falling standard condition (FSC), was similar to the first but with the ratio of falling and rising tones inverted with, 90% of the utterances having a falling F0 and the remaining 10% having a rising F0. The subjects in both conditions were first habituated to a 5-minute stream of utterances having the condition-specific F0 distribution while asked to perform an overt task that was designed to ensure that the participants were paying attention to the utterances. After habituation, they were tested in their judgments of prominence on a set of new utterances that also had falling or rising F0 trajectories during the last word.

Stimuli

Speech samples from the CAREGIVER Y2 FI corpus (Altosaar et al., 2010) were used in the study. The style of speech in CAREGIVER is enacted infant-directed speech (IDS) spoken in continuous Finnish, corresponding to a situation where a caregiver is talking to a child, and recorded in high quality in a noise-free anechoic room. The corpus was selected due to the availability of multiple sentences with a simple subject-verb-object (SVO) syntactic

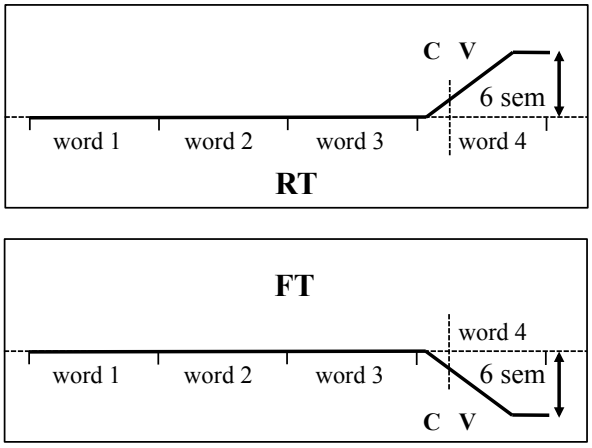

Figure 1: F0 trajectories for the two conditions. Top panel: contour for the rising trajectory (RT), Bottom panel: contour for the falling trajectory (FT). The dashed vertical line represents the vowel onset.

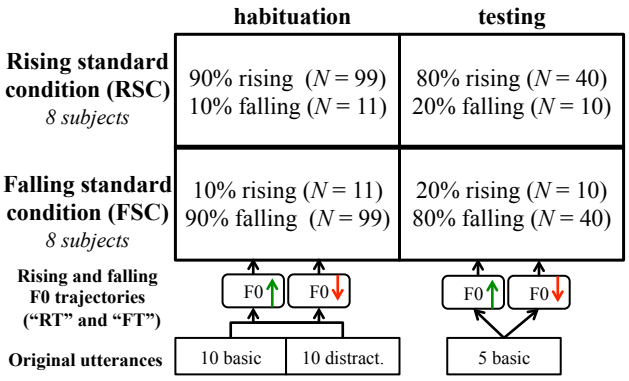

Figure 2: Overview of the experimental setup.

structure and slow speaking rate that allows easier manipulation of the F0 patterns in the sentences. After initial testing, 25 unique utterances from a female speaker (*Speaker 2*) were chosen for the experiments based on the overall naturalness of the stimuli after manipulating the F0 trajectories of the utterances (see below). Each stimulus consisted of a four-word SVO sentence with an average duration of three seconds.

Ten of the utterances were assigned as the main habituation stimuli and five utterances were used solely for the testing stage. The last ten utterances formed a so-called *distractor set* since, despite being grammatically correct, they had very unusual semantic structure (e.g., "*Vauva antaa kulmikkaan koiran*.", Eng: "*The baby gives the square dog*."). The distractors were used as targets in the overt task given to the listeners during habituation (see below).

For each of the 25 utterances, we generated two prosodic versions with either a falling F0 trajectory (FT) or rising F0 trajectory (RT) on the last word using the pitch manipulation functionality available in the Praat software (Boersma & Weenink, 2012). For each stimulus, the original pitch contour was first flattened and the F0 set to 185 Hz (approximately the average across the original stimuli), reflecting the most natural sounding pitch level for the speaker. The F0 trajectory was then modified for the fourth word ("target word") of each stimulus while keeping the rest of the contour flat (Figure 1). Since the primary stress of Finnish always falls on the first syllable of a word (e.g., Suomi & Ylitalo, 2002), the pitch excursion for the falling and rising trajectories was set to start right before the vowel onset of the first syllable of the fourth word (see, e.g., Hermes & Rump, 1994). To ensure consistency, the pitch excursion for all stimuli started 50-ms before the vowel onset and peaked 150-ms later, staying constant for the remaining part of the utterance (Figure 1). As the relation between pitch excursion size and prominence perception may vary, some studies reporting that a difference of 1.5 semitones (Rietveld & Gussenhoven, 1985) or even 4 semitones (Hart, 1981) is required for a perceptually noticeable difference, we preliminarily experimented with a number of different excursion sizes. After assessing excursion sizes between 2 and 8 semitones, we selected 6-semitone change from the flat F0 as the difference producing a clear and most natural-sounding perception of prominence across all stimuli. Thus, the modified trajectories during the target words had ±6 semitone excursion from the flat 185 Hz F0 of all previous words in the utterance (Figure 1). After resynthesizing the utterances with the modified F0 trajectories, all stimuli were amplitude normalized.

Participants

Sixteen native Finnish speakers (9 male, 7 female; average age 28 years) participated in the listening experiment. The test subjects were recruited from the personnel and students of Aalto University. All participants reported normal hearing. The subjects were randomly assigned to the two test conditions with 8 subjects per condition.

Experimental procedure

The listening experiment was conducted in a sound-isolated listening booth of the Acoustics Laboratory of the Aalto University. The habituation and testing software was run on a Mac mini with Matlab 2014b. The audio from the computer was fed through a Motu UltraLite-mk3 Hybrid into a pair of high-quality Sennheiser HD650 headphones.

Participants were given a brief description of the task by the experimenter and they were then asked to start the experiment. The experiment consisted of two parts: (i) a habituation and (ii) a testing stage (see Figure 2 for an overview). During the habituation stage, participants were asked to listen carefully to each utterance being played and press the spacebar whenever they heard a semantically incoherent sentence (the distractor). The role of the overt task was to ensure that the subjects engaged into holistic lexical and semantic processing of the stimuli during the habituation. The subjects were not given specific instructions regarding what counts as a semantically incoherent target, but were instructed to use their own judgment.

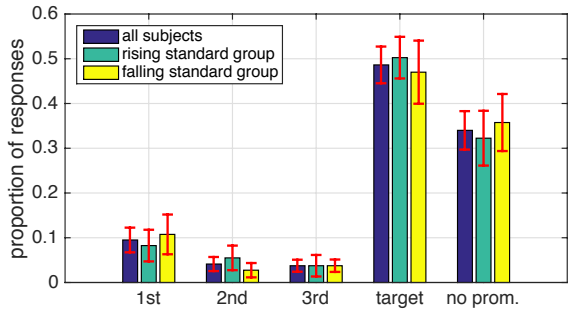

Figure 3: Proportion of responses to different words in the test sentences (relative position), including the no-prominence option. The error bars denote one standard error measured across test subjects (SE).

The only difference between the subject groups, i.e., the RSC and FSC conditions, was the distribution of the rising and falling F0 trajectories during the habituation and testing stage. In the case of RSC (FSC), 90% of the stimuli were RT (FT) ("standards") and 10% were FT (RT) ("deviants"). Ten semantically incoherent targets (distractor stimuli) were interleaved in the data and also followed the same 9:1 ratio. The ordering of the stimuli was randomized for each participant in a manner that each participant in each condition heard each lexically unique training utterance exactly the same number of times: nine times with RT (FT) and once with FT (RT). In addition, they heard each distractor once. The total duration of the habituation stage was 5 minutes, corresponding to a total of 110 utterances with a 500-ms silence interval between each utterance. To avoid repetition of the same utterance in a sequence, the stimuli were presented in blocks of 11 where each block had 10 unique training sentences and one distractor. There were no audible pauses between the blocks.

In the second part of the experiment, the subjects heard the test utterances one-by-one and, for each utterance, they were asked to grade the prominence level of only the single most prominent word on a nominal scale of 0 = no prominence, 1 = slight prominence, 2 = notable prominence. Subjects were allowed to hear each utterance only once in order to facilitate the capture of initial perceptual impressions and in order not to alter the perceived distribution of the pitch trajectories. The tasks of word selection and prominence grading were carried out using a graphical user interface (GUI) where a list of the spoken words was presented together with the prominence scale. Subjects used a mouse as the controller to make the selection. The stimuli distribution in the test stage was set to have 80% of standards (RT for RSC; FT for FSC) and 20% of deviants (FT for RSC; RT for FSC) in order to get more test samples for the deviants than what would be available from the original habituation distribution. The stimuli were presented in blocks of 5 comprising, for the case of RSC, 4 RT and 1 FT. There were 10 blocks of test stimuli, adding up to a total of 50 test tokens per subject. There were no distractor stimuli during the testing stage.

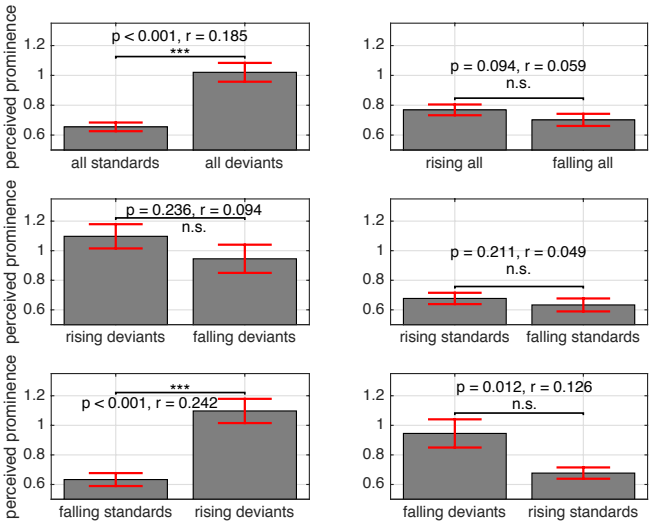

Figure 4: Means and SEs for the perceived prominence (0–2) levels for different stimulus types. Significance and effect sizes are reported using the Wilcoxon rank-sum test and using Bonferroni corrected significance level of $p < 0.0083$.

Results

We first verified how often the subjects labeled utterance-final words, the targets, as prominent instead of other words in the sentences. Figure 3 shows the proportion of responses to each of the four words (relative positions) in the test sentences and no-prominence responses across all subjects and for both sub-groups separately. As can be observed from the figure, the subjects primarily considered the manipulated target word as a prominent or not prominent with less than 18% of the responses marking one of the first three words in the utterances as prominent. In addition, the response strategies between the two subject groups are similar, suggesting that both types of pitch patterns were similarly strong attractors of prominence perception with respect to other competing words in each utterance.

Since the lexical content of the stimuli was fully independent of the pitch trajectory, each subject hearing exactly the same number of repetitions for each sentence, and each sentence occurring exactly the same number of times for both falling and rising pitch across both subject groups, we were able to pool the responses using the standard/deviant criterion, the absolute direction of the pitch pattern, or their combination. Figure 4 shows the full summary of perceived prominence levels across a number of different comparisons.

The results indicate that there is a main effect of condition with the less frequent pitch trajectories (deviants) perceived more prominent (M = 1.02, SD = 0.759) than the standard trajectory experienced during the habituation (M = 0.655, SD = 0.66) independently of the direction of the change. Although the effect is not large, it is highly significant ($p <$ 0.001, Z = 5.44, r = 0.185; Wilcoxon rank-sum test). In contrast, there are no observable differences between the perceived prominence of rising and falling pitch patterns when pooled across both groups, between rising and falling

deviants, or between rising and falling standards ($p > 0.0083$ after Bonferroni corrected for multiple comparisons; see Figure 4 for details). In other words, only the probability of the pitch trajectory during the habituation and testing has an impact on the overall prominence levels.

A closer analysis of both subject groups shows that there are indications for a reversal of preferences between the groups (Figure 4, bottom panel): Subjects habituated with frequent rising pitch consider words with a falling pitch marginally more prominent ($M = 0.945$, $SD = 0.815$ for falling and $M = 0.677$, $SD = 0.613$ for rising; $Z = 2.510$, $p = 0.012$, $r = 0.126$) whereas subjects habituated with predominant falling pitch consider words with a rising pitch significantly more prominent ($M = 1.097$, $SD = 0.695$ for rising versus $M = 0.633$, $SD = 0.705$ for falling; $Z = 4.847$, $p < 0.001$, $r = 0.242$). Although the effects are not large, the reversal of preferences between the two subject groups is clearly seen in the data.

Discussion and conclusions

The present findings suggest that the statistical distribution of prosodic cues can impact subjective perception of word prominence. This is in line with the earlier work that connects the idea of prominence to low-predictability events in the perceptual stream, i.e., "something standing out from a context" (Kakouros & Räsänen, in press; cf. Itti & Baldi, 2009) but contrasts with the idea of prosodic stress being conveyed with certain type of prosodic patterns such as rising or falling pitch contours (see, e.g., Hermes & Rump, 1994). Since there were no lexical or semantic differences between the low- and high-predictability targets, the present results also show that this type of expectation-based prosodic processing occurs in parallel to lexical processing.

From a cognitive point of view, a predictability-based system for prominence (i.e., attentional capture) would be much more flexible than one based on a fixed set of acoustic/prosodic feature detectors for prominent words. First of all, it enables "learning" of prominence perception from language experience, enabling a natural way for different languages to develop strategies for conveying prominence when constrained by the simultaneous production of phonemic contrasts of the language. Similarly, a predictability-based system allows shorter time-scale adaptation to the ongoing communicative situation where factors such as communication channel (is speech partially masked by noise?) or talker-specific idiosyncrasies (acoustic and linguistic characteristics of a specific talker) can lead to very different acoustic prosodic outcomes than what can be characterized at a language-general level. Finally, the predictability framework integrates naturally to the work on statistical learning and information theoretic models at different levels of linguistic analysis (see, e.g., Jurafsky et al., 2001; Frank & Jaeger, 2008; Aylett and Turk, 2004) and language learning (see, e.g., Saffran, Aslin, & Newport, 1996), suggesting that similar basic mechanisms for capturing statistical regularities in the sensory input may be responsible for phenomena at multiple different levels and domains of cognitive behavior.

However, the present findings are preliminary and should be investigated further in a number of additional experiments. For instance, the present setup used an arbitrarily chosen ratio of 1 deviant to 9 standards during habituation, revealing that the subjects are sensitive to such a difference. It is unclear whether the magnitude of apparent prominence is related to the probabilities of the tokens or whether the mechanism is more binary in nature. In addition, the fundamental frequency is only one of many cues to word prominence (see, e.g., Fry, 1955; 1958; Lieberman, 1960; Ortega-Llebaria & Prieto, 2010) and factors such as word position and durational cues also play a role (see, e.g., Luchkina & Cole, 2014). Since we used pre-recorded sentences from a corpora originally designed for other purposes, the speech is typical continuous speech in the sense that there is a high likelihood to have stress on sentence-final words. Despite controlling for pitch and energy, subtle cues such as fine-grained timing in syllabic structure may still be present, here seen as an inherent bias to perceive standard targets also as slightly prominent ($M = 0.655$, $SD = 0.660$, on the scale of 0–2). However, these cues were exactly the same for both of our test groups and cannot affect the group differences.

In all, the current findings provide initial behavioral support for the hypothesis that prominence and unpredictability of the acoustic prosodic features are connected. However, more work is needed in order to confirm this finding and to understand the characteristics and limits of the probabilistic framework in the perception of speech prosody. This also includes extension of the study for other languages than Finnish.

Acknowledgments

This study was funded by the Academy of Finland in the project "Computational modeling of language acquisition", Aalto ELEC Doctoral School, and Nokia Foundation.

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-2010)* (pp. 1062–1068).

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*, 31–56.

Aylett, M., & Turk, A. (2006). Language redundancy predicts syllabic duration and the spectral characteristics of vocalic syllable nuclei. *Journal of the Acoustical Society of America*, *119*, 3048–3058.

Boersma, P. & Weenink, D. (2012). Praat: doing phonetics by computer [Computer program]. Version 5.3.13, retrieved from http://www.praat.org/

Calhoun, S. (2010). The centrality of metrical structure in signaling information structure: A probabilistic perspective. *Language*, *86*, 1–42.

Campbell, N. (1995). Loudness, spectral tilt, and perceived prominence in dialogues. *Proceedings of the 13th International Congress of Phonetic Sciences (ICPhS-1995)* (pp. 676–679).

Campbell, N., & Beckman, M. E. (1997). Stress, prominence, and spectral tilt. In A. Botinis, G. Kouroupetroglou, & G. Carayiannis (Eds.), *Intonation: Theory, Models, and Applications (Proceedings of an ESCA Workshop)* (pp. 67–70).

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

Cutler, A., & Foss, D. J. (1977). On the role of sentence stress in sentence processing. *Language and Speech*, *20*, 1–10.

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

Frank, A., & Jaeger, T. F. (2008). Speaking rationally: Uniform information density as an optimal strategy for language production. *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 933–938). Austin, TX: Cognitive Science Society.

Fry, D. B. (1955). Duration and intensity as physical correlates of linguistic stress. *Journal of the Acoustical Society of America*, *27*, 765–768.

Fry, D. B. (1958). Experiments in the perception of stress. *Language and Speech*, *1*, 126–152.

Hermes, D. J., & Rump, H. H. (1994). Perception of prominence in speech intonation induced by rising and falling pitch movements. *The Journal of the Acoustical Society of America*, *96*, 83–92.

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

Jurafsky, D. (1996). A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science*, *20*, 137–194.

Jurafsky, D., Bell, A., Gregory, M., & Raymond, W. D. (2001). Probabilistic relations between words: Evidence from reduction in lexical production. *Typological studies in language*, *45*, 229–254.

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

Kakouros, S., & Räsänen, O. (2015a). Analyzing the Predictability of Lexeme-specific Prosodic Features as a Cue to Sentence Prominence. *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 1039–1044). Austin, TX: Cognitive Science Society.

Kakouros, S., & Räsänen, O. (2015b). Automatic Detection of Sentence Prominence in Speech Using Predictability of Word-Level Acoustic Features. *Proceedings of the 16th Annual Conference of the International Speech Communication Association* (pp. 568–572). Dresden, Germany: International Speech Communication Association.

Kakouros, S., & Räsänen, O. (in press). Perception of sentence stress in speech correlates with the temporal unpredictability of prosodic features. *Cognitive Science,* accepted for publication.

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

Luchkina, T., & Cole, J. (2014). Structural and prosodic correlates of prominence in free word order language discourse. *Proceedings of Speech Prosody (SP-2014)*.

Maier, A. K., Hönig, F., Zeißler, V., Batliner, A., Körner, E., Yamanaka, N., & Nöth, E. (2009). A language-independent feature set for the automatic evaluation of prosody. *Proceedings of Interspeech* (pp. 600–603).

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

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

Rietveld, A. C. M., & Gussenhoven, C. (1985). On the relation between pitch excursion size and prominence. *Journal of Phonetics*, *13*, 299–308.

Rosenberg, A., Cooper, E. L., Levitan, R., & Hirschberg, J. B. (2012). Cross-language prominence detection. *Proceedings of Speech Prosody*.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*, 1926–1928.

Shattuck-Hufnagel, S., & Turk, A. E. (1996). A prosody tutorial for investigators of auditory sentence processing. *Journal of psycholinguistic research*, *25*, 193–247.

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*, 2471–2485.

Suomi, K., & Ylitalo, R. (2002). Word stress and duration in Finnish. In C. Dunger, B. Granström, D. House & A. M. Öster (Eds.), *Proceedings of the Swedish Phonetics Conference (Fonetik-2002)* (pp. 73–76).

t Hart, J. (1981). Differential sensitivity to pitch distance, particularly in speech. *The Journal of the Acoustical Society of America*, *69*, 811–821.

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

Vaissière, J. (1983). Language-independent prosodic features. In A. Cutler & D. R. Ladd (Eds.), *Prosody: Models and measurements* (pp. 53–66). Springer Berlin Heidelberg.

Wagner, P., Origlia, A., Avesani, C., Christodoulides, G., Cutugno, F., D'Imperio, M., ... & Vainio, M. (2015). Different Parts of the Same Elephant: a Roadmap to Disentangle and Connect Different Perspectives on Prosodic Prominence. In The Scottish Consortium for ICPhS 2015 (Ed.), *Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS-2015)*. Glasgow, UK: the University of Glasgow.