Perception of sentence stress in speech correlates with the temporal unpredictability of prosodic features

Sofoklis Kakouros[a] & Okko Räsänen[a]

[a]*Department of Signal Processing and Acoustics, Aalto University*

Keywords: Stress perception; Attention; Statistical learning; Sentence stress; Prosody; Statistical modeling; Stimulus predictability

Correspondence should be sent to Sofoklis Kakouros, Department of Signal Processing and Acoustics, Aalto University, Otakaari 5, 02150 Espoo, Finland. E-mail: sofoklis.kakouros@aalto.fi

Abstract

Numerous studies have examined the acoustic correlates of sentential stress and its underlying linguistic functionality. However, the mechanism that connects stress cues to the listener's attentional processing has remained unclear. Also, the learnability versus innateness of stress perception has not been widely discussed. In this work, we introduce a novel perspective to the study of sentential stress and put forward the hypothesis that perceived sentence stress in speech is related to the unpredictability of prosodic features, thereby capturing the attention of the listener. As predictability is based on the statistical structure of the speech input, the hypothesis also suggests that stress perception is a result of general statistical learning mechanisms. In order to study this idea, computational simulations are performed where temporal prosodic trajectories are modeled with an n-gram model. Probabilities of the feature trajectories are subsequently evaluated on a set of novel utterances and compared to human perception of stress. The results show that the low-probability regions of F0 and energy trajectories are strongly correlated with stress perception, giving support to the idea that attention and unpredictability of sensory stimulus are mutually connected.

1. Introduction

Prosody is a very important characteristic of spoken language that can be seen as defining the underlying segmental structure of an utterance or as a set of acoustic parameters in continuous speech. In this work, we use the latter definition and view prosody as a time series of acoustic-related information. Typical parameters include (i) the pitch (fundamental frequency), (ii) loudness (intensity), (iii) duration (see Cutler, Oahan, & van Donselaar, 1997 for a review), and (iv) spectral tilt (see Sluijter & van Heuven, 1996). The prosodic encoding taking place during speech production is largely determined by the speaker's choices and contains non-linguistic information that is not available in the written counterpart of the communicated message. This also means that different speakers may apply their own prosodic modifications on the acoustic parameters of the same sentence or phrase in order to convey certain communicative intent which is then reflected in the perceptual processing of the target listener. Sentence stress is one type of prosodic specification where one or more words within a sentence receive special emphasis. Similar phenomena, on different domains, are often found in the literature under the terms of sentence, phrasal and lexical stress or focus (see, e.g., Cutler & Foss, 1977; Cutler et al., 1997; Cutler, 2005; Werner & Keller, 1994; Gussenhoven, 2011). As the terminology can be ambiguous, in the current work the term sentence stress will be used to denote the most prominent word or words within a sentence.

The majority of the earlier work on stress has focused on (i) the linguistic and phonetic level, where the aim has been to identify the acoustic correlates of stress and their function in language perception (see, e.g., Chrabaszcz, Winn, Lin, & Idsardi, 2014, for a cross-linguistic study; Campbell, 1995, for American English) and (ii) computational models attempting to automatically detect stress by studying a number of acoustic and linguistic cues (see, e.g.,

Mishra, Sridhar, & Conkie, 2012; Wang & Narayanan, 2007, for American English). There is also a growing body of literature focused on understanding how prosodic correlates of stress are reflected in brain activity, such as in terms of event-related potentials (ERPs) (see, e.g., Pannekamp, Toepel, Alter, Hahne, & Friederici, 2005; Steinhauer, 2003).

The general finding from the existing research on speech production is that the acoustic realization of stress is typically manifested as changes in the fundamental frequency (F0), intensity, or duration of the syllables or words compared to their unstressed variants (e.g., Werner and Keller, 1994) or as changes in spectral tilt in the vowel nuclei between stressed and unstressed syllables (e.g., Sluijter & van Heuven, 1996). In most of the world's languages, at least a subset of these features can be controlled in speech production relatively independently of the lexical content of the message. From the listener's point of view, however, the relationship between sentence stress and the attentional mechanisms modulating the perceptual processing of the speech input has not been systematically investigated. Also, little attention has been given to the question whether perception of stress is based on innate sensitivity to specific acoustic cues or whether it is learned from language experience.

In this work, we propose that the perception of stress in speech is related to the temporal unpredictability of the acoustic features. Earlier work has suggested a connection between the expectations of F0 trajectories and stress perception (see, e.g., Terken, 1991) but these findings have not been formulated into an explicit model. Our hypothesis is that the talker is capable of focusing the listener's attention on desired parts of the speech stream by manipulating the predictability of the prosodic cues. As for the listener, perceptual attention is driven by the unpredictability of the speech input, allocating processing resources to those aspects of the input that are the most informative (i.e., parts that differ the most from the internal expectations of the

situation). This means that stress perception may not be necessarily based on specific configurations of acoustic features such as high pitch or energy, but based on the deviation of these features from the listener's expectations based on the earlier language experience and the current communicative context (see also Aylett & Turk, 2004). The unpredictability hypothesis converges with the visual neuroscience research where it has been suggested that the strongest attractors of visual attention are stimuli that stand out from their neighbors in space or time (Itti & Baldi, 2005, 2009). In order to narrow the scope of the current study, we focus solely on the study of sentential stress. However, there is no reason to assume that the same type of mechanism would not be operational at different levels of representation.

We test our hypothesis by first investigating human stress perception in a listening test using both native and non-native listeners and then comparing the behavioral findings to the output of an unsupervised computational model of statistical learning. During a learning stage, the model learns the typical temporal evolution of energy, F0, spectral tilt, and full spectrum of speech but is ignorant of the relationship between the features and the concurrent presence or absence of sentential stress. After learning, the model can be used to compute probabilities for the feature trajectories in new sentences. When the output of the algorithm is compared to human perception of stress in the same set of sentences, the points of high unpredictability are observed to be highly correlated with the words that are perceived as stressed by human listeners.

This paper is organized as follows: First, earlier work and the most important findings in speech perception research on sentential and lexical stress are reviewed, followed by the current approaches in computational models for stress detection. In section 2, we attempt to connect stress perception, attentional orientation, and the statistical learning paradigm under a single framework. Section 3 describes the data collection method and analysis, while section 4

describes the statistical model, followed by the experimental results. Finally, a discussion and conclusions are presented in the last section.

1.1. Perception of stress and its prosodic correlates

Speech is a particularly rich signal containing lexical and grammatical information (what is said), prosodic (how it is said), and other, speaker dependent information (such as the identity and emotional state of the speaker). More specifically, prosody is related to speech features whose domain is larger than one phonetic segment, concerning syllables, words, phrases, sentences, and even longer utterances—also known as *supra-segmental features* (Werner & Keller, 1994; see also Lehiste, 1970). In this context, stress can be defined as an accentuation of syllables within words or of words within sentences (Cutler, 2005). When specific words are emphasized in the context of entire utterances, the phenomenon is referred to as sentence prominence or sentence stress.

Stress in speech is typically analyzed in terms of its acoustic correlates. Most common cues associated with sentence and word stress are the changes in F0, intensity, duration, and spectral slope (or tilt) of speech (see, e.g., Bolinger, 1964; Fry, 1955; Lieberman, 1960; Terken, 1991; Kochanski, Grabe, Coleman, & Rosner, 2005; Sluijter & van Heuven, 1996; Chaolei, Liu, & Shanhong, 2007; Ortega-Llebaria & Prieto, 2010; Campbell, 1995; Campbell & Beckman, 1997). For example, Terken (1991) studied the effect of F0 on prominence using artificial words, concluding that F0 plays an important role in predicting stress. However, he also noted that its function in predicting stress is more complex than a simple manifestation of a local maximum or change in F0 and that both global and local properties of the F0 contours must be taken into consideration. According to the study of Sluijter and van Heuven in Dutch (1996), stressed syllables seem to be perceived as louder and more prominent than their unstressed counterparts

due to an increased energy level at the higher frequencies. Accordingly, spectral tilt could be another important acoustic correlate of stress as it is indicative of the energy distribution between the low and high frequencies. However, Ortega-Llebaria and Prieto (2010) note that the relationship between spectral tilt and syllable stress has not been fully established, since studies on different languages have reached conflicting findings on the topic (see, e.g., Sluijter & van Heuven, 1996, for Dutch; Campbell & Beckman, 1997, for American English). Finally, loudness and duration are known to be important contributors, at least in the perception of word stress, as prominent syllables are typically louder and longer than the non-stressed syllables (see, e.g., Kochanski et al., 2005, for British English). Overall, the acoustic correlates of stress seem to be descriptive of stress in a number of languages though the language-specific realization may vary and all or a subset of the correlates can be used in order to convey stress (see, e.g., Malisz & Wagner, 2012, for Polish; Tamburini & Wagner, 2007, for German; Tamburini & Caini, for American English; Ortega-Llebaria, 2006, for Spanish; Eriksson, Barbosa & Akesson, 2013, for Swedish).

A number of other studies have examined stress with respect to its function in perceptual processing. In this regard, sentence stress seems to have effects on the parsing of information and syntactic structure from utterances (see, e.g., Shattuck-Hufnagel & Turk, 1996; Cutler et al., 1997; Gussenhoven, 2011). For instance, Cutler and Foss (1977) measured the reaction times (RT) to word-initial phoneme targets on content and function words in sentence contexts. They found that RTs were shorter for stressed words independently of their syntactic function (see also Shields, McHugh, & Martin, 1974), suggesting that stress enabled rapid and efficient recognition of speech patterns during sentence processing. As more salient words seem to receive more processing through attention, accenting new information in simple comprehension tasks results

in shorter response times (Bock & Mazzella, 1983; Birch & Clifton, 1995). There also seem to be both language-universal and language-specific relationships between prosodic prominence and syntactic structure (Cutler et al., 1997). For instance, stress may not be necessarily reflected by differences in F0, as in English, but instead by a difference in word order (see also Ladd 2008). Nonetheless, Endress and Hauser (2010) argue that there are universally available prosodic characteristics that assist in finding words in connected speech despite the differing sound structures across languages.

Finally, a number of studies have investigated the prosodic patterns which take place before and during stress perception, focusing on the predictive and probabilistic nature of the stress occurrence. For instance, intonation patterns may help listeners predict accents (see e.g. Cutler & Foss, 1977; Cutler & Darwin, 1981; Calhoun, 2007). The work of Calhoun (2007) also suggests that stress is likely to be perceived for words that are acoustically and structurally more prominent than anticipated based on their syntactic, semantic and discourse properties. This idea implies a strong probabilistic connection of the alignment of words to prosodic structure. In this regard, prosody can be seen as organization of information within an utterance by highlighting or toning down parts of the utterance (Calhoun, 2010). Other work has also examined the rhythmic structure of speech as a predictor for an upcoming point of stress (see, e.g., Shields et al., 1974; see also Ladd, 2008).

1.2. Computational models of stress detection

In contrast to the linguistic research, computational models of sentential stress are typically focused on automatic detection of stress from speech signals for analysis or application-specific purposes and the proposed methods can be roughly divided into supervised and unsupervised. Supervised methods typically involve a data-intensive approach where human annotators provide

labeling for a set of training data after which a statistical model is taught the link between stressed units and the acoustic features of speech (see, e.g., Minematsu, Kobashikawa, Hirose, & Erickson, 2002; Imoto, Tsubota, Raux, Kawahara, & Dantsuji, 2002; Moubayed, Ananthakrishnan, & Enflo, 2010; Lai, Chen, Chu, Zhao, & Hu, 2006; Li, Zhang, Li, Lo, & Meng, 2011; Chaolei et al., 2007). Other supervised methods use acoustic and linguistic features together (e.g., lexical, part of speech tags, information structure) for stress detection (see, e.g., Sridhar, Nenkova, Narayanan, & Jurafsky, 2008; Calhoun, 2007) or operate purely on linguistic features (see, e.g., Hirschberg, 1993). Even though supervised methods perform well on inputs similar to the training data, their usage is restricted to languages with sufficient labeled data. More importantly, they are not models for human stress perception since the training process involves the use of input labeling —information that is not available for human learners.

Instead of using *a priori* linguistic knowledge, unsupervised methods extract acoustic features directly from speech and use them in order to calculate prominence levels (see, e.g., Wang & Narayanan, 2007; Kalinli & Narayanan, 2009; Tamburini & Caini, 2005; Tamburini, 2003; Mehrabani, Mishra, & Conkie, 2013; Imoto, Dantsuji, & Kawahara, 2000; Rosenberg & Hirschberg, 2009). In these models, the criteria for determining stress from the signal are typically prominence scores calculated over automatically extracted syllabic nuclei using different feature combinations. For example, Tamburini (2003) used a prominence function to compute scores over syllabic nuclei utilizing energy, nucleus duration, and event amplitude as parameters. Prominent syllables were then selected by finding the maximum syllable score from its two neighboring syllables and evaluated relative to manually annotated prominence with high rates of agreement between the two. The current work also follows the unsupervised approach by

modeling the temporal evolution of the prosodic features purely on the basis of the speech input and without assuming any *a priori* linguistic knowledge during the learning stage.

2. Stress, stimulus-driven attention, and statistical learning

As the role of sentential stress is to emphasize specific words in the spoken utterance, it is important to consider the cognitive mechanisms that actually modulate the listener's behavior relative to the stress-related acoustic cues. We believe that stimulus-driven attentional modulation is central to stress perception. In the current work, the concept of attention will refer solely to stimulus-induced, bottom-up switching of perceptual attention. More specifically, we claim that stress is an aspect of prosody whose role in perception is analogous to perceptual orientation (Sokolov, 1963) and that the perception of stress is based on statistical learning mechanisms. In models of visual perception, the traditional approach is to divide processing into basic preattentive analysis that provides an array of functional perceptual units and attentive processing that selects a subset of the units for more resource-demanding but detailed, possibly conscious, analysis (see, e.g., Folk, Remington, & Johnston, 1992, and references therein). One possibility is that auditory perception has a similar capability to focus on specific temporal segments of the signal for a more detailed analysis, possibly leading to enhanced processing of the segments in comparison to standard processing. For instance, focus on specific words may speed up their recognition or improve word learning.

Typically, stimulus-driven attention is connected to the concept of saliency (in the visual domain) or novelty (in the auditory domain) (Itti & Baldi, 2009; Ranganath & Rainer, 2003). For instance, in the visual domain, a black dot appearing on a white background would immediately draw our attention. Intuitively, saliency seems to be associated with stimuli that have extreme physical properties, such as very loud noise, extreme brightness, or, in the case of speech, high

energy or pitch. However, beyond some rare exceptions, the typical attention-capturing stimuli are not extreme on any absolute scale but simply stand out or are unpredictable in their current context (Itti & Baldi, 2009; Corbetta & Shulman, 2002). In the case of the black dot on the white screen, consistently recurring appearance and disappearance of the dot would no longer draw attention as the process would quickly become fully predictable.

If saliency is driven by unpredictability, there has to be some type of model for the current overall sensory input ("context") against which the individual parts of the input can be compared. At a computational level, such a model can be seen as a statistical description of the context (e.g., a visual scene or spoken utterance), including characterization of the present entities and events and their dependencies across time and space. The more a stimulus differs from the description encoded in the context model, the more it is perceived as salient (cf., Itti & Baldi, 2009). The same idea is also mathematically formalized in information theory where the information value of an event is inversely proportional to the probability of observing the event (Shannon, 1948). Importantly, the concept of context-dependent saliency and self-information both imply that at least some sort of *learning* or *representational persistence* (memory) is required—a statistical description of the current context needs to be actively represented in the system in order to detect violations from the expectations.

It is now well known that the human brain has evolved to learn statistical regularities of the environment. For example, organization of the early sensory cortices is driven by properties of the incoming sensory stimuli (Blakemore & Cooper, 1970; Sur, Garraghty, & Roe, 1988). Moreover, infants are known to be sensitive to the distribution of acoustic features in speech across acoustic and temporal domains (Kuh, Williams, Lacerda, Stevens, & Lindblom, 1992; Maye, Werker, & Gerken, 2002; Saffran, Aslin, & Newport, 1996; Saffran, Newport, & Aslin,

1996; Saffran, 2001) and also to the cross-modal connections between auditory and visual domains (e.g., Smith & Yu, 2008; Teinonen, Aslin, Alku, & Csibra, 2008). It has been established that the statistical learning is not limited to language, but is observed across the perceptual domains (Saffran, Johnson, Aslin, & Newport, 1999; Baldwin, Andersson, Saffran, & Meyer, 2008; see also Romberg & Saffran, 2010, for a review).

Following the behavioral findings, computational models have been used to investigate what kinds of statistical regularities are available in speech and what kind of learning mechanisms are required for their acquisition. Previous studies have provided evidence that bootstrapping of word learning is possible without existing knowledge of phonetic categories or segmental structure of words (Räsänen, 2011), that distributional properties of phonetic categories can be modeled in an unsupervised manner (e.g., Feldman, Griffths, & Morgan, 2009; Lake, Vallabha, & McClelland, 2009), and that a simple statistical learning mechanism is capable of explaining behavioral findings on rule-like learning (Laakso & Calvo, 2011; Räsänen & Rasilo, 2012; see also Räsänen, 2012 for a review).

In general, there is accumulating evidence pointing towards a set of general statistical learning mechanisms capable not only of differentiating structure from randomness, but also of judging the strength (relative probability) of different statistical patterns at different levels of representation. A system supported by such learning would also provide a solid basis for forming prosodic expectations in different contexts and thereby also enable the detection of statistical deviation and attentional capture by unpredictable sensory events. Existing data also support this idea. In ERP studies, the widely studied mismatch negativity signal reveals clear pre-attentive processing of unpredictable deviants, and the effect persists even during top-down attempts to ignore the stimuli (see, e.g., Sussman, Ritter, & Vaughan, 1998). Astheimer and Sanders (2009,

2011) have shown that unpredictability of word onsets in speech strongly correlates with the listener's attention to these segments and that unpredictable words elicit larger ERP signals (N1) than their more predictable counterparts. Finally, Itti and Baldi (2009) have shown that the statistical unpredictability of visual regions in videos, measured in terms of a Bayesian model, correlate better with human eye-fixation behavior than image features or saliency maps describing the overall level of change and/or contrast in the picture.

Following the work of Itti and Baldi (2009) on visual attention, we investigate whether the surprisal in prosodic trajectories correlates with the perception of prominence in speech. As a talker is able to control the degree of temporal predictability in an utterance through the use of supra-segmental cues, the talker can also mark parts of the signal as "novel", thereby inducing the altered cognitive processing for those parts of the message (see, e.g., Ranganath & Rainer, 2003). If this were the case, then sentence stress would not be simply defined on the basis of the presence or absence of specific pre-defined acoustic cues, but on the overall predictability of the acoustic correlates of prosody in the given communicative context. Therefore, stress perception would be an outcome of general statistical learning mechanisms. In contrast to Calhoun (2007), we do not assume syntactic or semantic parsing in the model but simply assume that the learner is able to learn the statistical properties of acoustic speech features, making the model plausible for stress perception during early stages of language acquisition.

3. Data collection and analysis

3.1. Data collection

3.1.1. Participants

A total of twenty test subjects (11 male, 9 female, age range 20–61 years with a median of 30 years) participated in the listening experiment. The participants were recruited from the

personnel and students of Aalto University and University of Helsinki, Finland. Fourteen of the participants were L1 (first language) Finnish speakers and six of the participants were L1 UK English speakers. English was the L2 (second language) of all Finnish listeners, all of whom classed themselves as professional-level English users. Six of the L1 Finnish subjects also took the LexTALE proficiency test on English as a post-hoc control procedure, achieving an average score of 92.08/100 in the test, corresponding approximately to C1 & C2 in the Common European Framework (CEF) language proficiency levels (Lemhöfer & Broersma, 2012). Five of the six L1 English speakers also took the test afterwards with an average score of 98.25/100. All participants reported normal hearing.

3.1.2. Apparatus

The listening experiment was conducted in a sound-isolated listening booth at Aalto University. The data collection software was run on a Mac mini with Matlab 2013a. The audio from the computer was fed through a Motu UltraLitemk3 Hybrid into a pair of high-quality Sennheiser HD650 headphones.

3.1.3. Speech stimuli

The CAREGIVER Y2 UK corpus (Altosaar et al., 2010) was used in the study. The style of speech in CAREGIVER is enacted infant-directed speech (IDS) in UK English, corresponding to a situation where a caregiver is talking to a child in a scene with jointly-attended objects and events, but recorded in high-quality in a noise-free anechoic room. The corpus was originally designed for early word learning studies and therefore the words corresponding to the context-related objects and events are referred to as keywords in the corpus. During the recording, talker text prompts were paired with visual pictures of the keyword objects together with a picture of

the infant that was being talked to. All talkers either were parents themselves or had other experience with young infants.

In addition to a set of 50 unique keywords, there are a number of verbs and function words used in the surrounding carrier sentences of the corpus, yielding a total vocabulary of 80 words. The sentences were generated by random sampling from the pool of keywords but by ensuring the grammatical correctness. The talkers were not separately instructed on the use of prosody, but they were simply asked to read the text prompts as they would talk to their own child (see Altosaar et al., 2010, for details). The corpus also contains orthographic transcriptions corresponding to each utterance with time-aligned information at the word level.

Overall, the "main talker section" of CAREGIVER contains 2397 sentences. A subset of 300 unique utterances were chosen for the listening tests from one male and one female talker (*Speakers 3* and *4*), yielding a total of 600 sentences. All single-word sentences were excluded from the data, and there were on average 5.9 words (2.8 seconds) per sentence. This set of utterances is referred to as the test set, as it was also used to probe the performance of the statistical prosodic model. To train the statistical model, 2000 non-test sentences per talker were used that are referred to as the training set.

3.1.4. Listening test procedure

Participants were initially given a brief description of the annotation task by the experimenter and were asked to "listen and mark zero or more words which are perceived to be stressed/prominent in each utterance". They were then seated inside a sound-isolated booth in front of a screen where they were asked to familiarize themselves with the annotation tool and

start the annotation procedure. On average, the task took approximately 1.5 hours for a listener to complete.

The annotation tool consisted of a graphical user interface (GUI) developed and run in Matlab 2013a (see Fig. 1). The GUI played each utterance through headphones, displayed the list of spoken words in a temporally ordered list, and then prompted the user to choose the words that were perceived as stressed using a computer mouse as the controller. For each utterance, the listener could mark any of the given words as stressed and could also listen to each sentence as many times as they wished. Data from nine L1 Finnish subjects and all six L1 English subjects were first collected by showing the audio waveform of each utterance in the annotation tool (see Fig. 1). To assure that the visual appearance of the waveform was not biasing the stress markings in favor of high-amplitude words, data were collected from an additional set of five L1 Finnish subjects using an otherwise identical procedure but without the waveform on the screen.

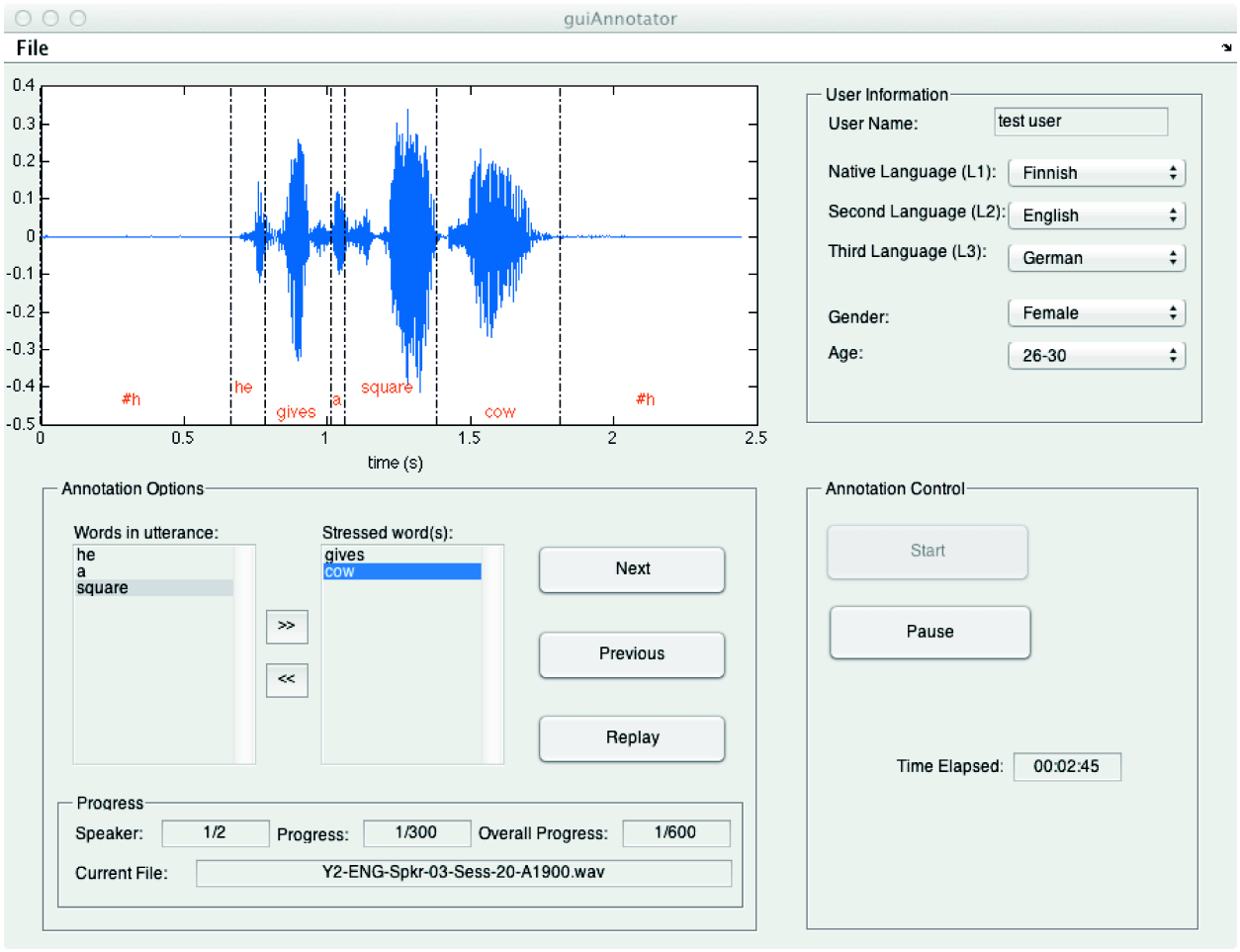

Fig. 1. Graphical user interface for data collection.

3.2. Listening test data analysis

3.2.1. Agreement-rate measures

The Fleiss kappa statistic (Fleiss, 1971) was used as the primary measure of the level of agreement across all listeners and between the listeners and the computational model (see section 4). In essence, Fleiss kappa measures the degree of agreement between two or more annotators on a nominal scale of $\kappa \in [-1,1]$. In the current work, all individual words occurring in the test set were considered as targets for a binary stress decision. Fleiss kappa takes into account the underlying distribution of the ratings, yielding $\kappa = 0$ if the number of agreements is equal to what

is expected based on chance-level co-occurrences in the data and $\kappa = 1$ if all annotators agree on all rated items.

Note that the Fleiss kappa computed over the entire pool of annotators is not the same as the arithmetic mean of the pair-wise agreement rates across all possible pairs of annotators. Thus, the agreement rate between the computational model and the human listeners had to be computed in a pair-wise manner with each individual listener (see section 5.1) or with respect to the annotators' majority decision on the most likely stressed word on a sentence level (see section 5.2). Therefore, both pooled overall and the mean pair-wise kappa were calculated from the data.

3.2.2. Data Analysis

Significance levels from all statistical tests in this sub-section are reported using the Mann-Whitney U test, since the data are not necessarily normally distributed. All tests were also carried out using the t-test, but because no qualitative changes in the results were observed, t-values are not reported.

The first step was to test whether there is an effect of L1 language in the inter-annotator agreement rates. Therefore the pair-wise Fleiss kappa scores between English listeners were compared to the agreement rates across each possible pair of a Finnish and an English listener. This analysis included only those nine L1 Finnish subjects who saw the signal waveform during the annotation procedure similarly to the six L1 English listeners. The mean pair-wise agreement was $\kappa_{EN} = 0.45$ (± 0.07 standard deviations) among the English listeners, $\kappa_{FI} = 0.40$ (± 0.11) among the Finnish listeners, and $\kappa_{EN\text{-}FI} = 0.42$ (± 0.10) between English and Finnish listeners. The difference between κ_{EN} and $\kappa_{EN\text{-}FI}$ was not significant ($p = 0.2784$) and neither was the difference

between κ_{EN} and κ_{FI} (p = 0.0705). In total, the analysis shows that there were no observable differences between L1 English and L1 Finnish listeners in the task.

In addition, the mean agreement ($\kappa_{FI\text{-}FI/nowave}$ = 0.42 ±0.11) between pairs of Finnish listeners, where one subject of the pair saw the signal waveform and the other did not, was not different from the mean agreement between the Finnish listeners who all saw the waveform during the annotation process (p = 0.3742). Similarly, the agreement across all pairs of English listeners with the visual waveform and Finnish listeners without the waveform did not differ from the mean agreement between the L1 English listeners only ($\kappa_{EN\text{-}FI/nowave}$ = 0.42 ±0.09 versus κ_{EN} = 0.45 ±0.07; p = 0.3795). Since neither the L1 of the listeners or the presence of the visual cues had an impact on the prominence perception, all twenty listeners were pooled together for the remaining analyses.

The overall Fleiss kappa across all 20 annotators was κ = 0.42, which translates into mean agreement rate of 85.0% for individual word tokens. On average, a total of 25.45% (±6.2%) of all words in the data were considered as stressed (approximately 1.5 stressed words per utterance). As for the pair-wise agreements between annotators, the mean agreement was also κ = 0.42 (± 0.10) with a minimum of κ = 0.19 and a maximum of κ = 0.65, reflecting a notable variation between the listeners in the task. Table 1 shows the inter-annotator agreement rates with respect to the male and female talker and for the male and female listeners.

Table 1. Fleiss kappa agreement-rates for gender-specific subsets in both listeners and talkers. Top half: Agreement rates computed across each entire subset. Bottom half: Means and standard deviations across all possible pairs of listeners in each subset.

Entire subset	Female talker	Male talker	Both talkers
Female listeners	0.48	0.34	0.41
Male listeners	0.49	0.38	0.43
All listeners	0.49	0.36	**0.42**
Mean pair-wise	**Female talker**	**Male talker**	**Both talkers**
Female listeners	0.47 (±0.07)	0.31 (±0.15)	0.40 (±0.10)
Male listeners	0.50 (±0.07)	0.36 (±0.10)	0.43 (±0.09)
All listeners	0.47 (±0.08)	0.34 (±0.13)	**0.42 (±0.10)**

As seen from Table 1, the mean pair-wise agreement on stress words is much higher for the female talker than the male talker (κ_{FT} = 0.47 vs κ_{MT} = 0.34, $p < 0.001$). This was also reflected in the informal comments made by the listeners after the listening task where several subjects found the female talker's prosodic patterns more prominent. Moreover, the higher agreement on the female talker is also significant when looking only at the female listeners (κ_{FL-FT} = 0.47 vs κ_{FL-MT} = 0.31, p = 0.0028) or male listeners (κ_{ML-FT} = 0.50 vs κ_{ML-MT} = 0.36, $p <$ 0.001) separately. Interestingly, the male listeners have significantly higher inter-annotator agreement rate on the male talker than the female listeners on the same talker (κ_{ML-MT} = 0.36 vs κ_{FL-MT} = 0.31, p = 0.0397) while the effect of listener's gender on the female talker (κ_{ML-FT} = 0.50 vs κ_{FL-FT} = 0.47, p = 0.1394) or both talkers together are not significant (κ_{ML-BT} = 0.43 vs κ_{FL-BT} = 0.40, p = 0.1740). In general, the current data from only two talkers is not sufficient to conclude whether there is some gender-specific effect in the stress patterns, but it shows that there are clear talker specific differences. Also, the reason why male listeners annotate the present male talker more consistently is not currently understood.

Finally, the relationship between word position and stress perception was investigated as stressed words are known to occur more often at the end of the utterances (see, e.g., Fernald & Mazzie, 1991). A total of 11.8% of the stress markings (N = 2140) were found to be on the first word of the utterances, 30.5% (N = 5542) on the last word, and the remaining 57.7% (N = 10480) were spread across the remaining positions in the utterances, in line with the earlier data (see, e.g., Fernald & Mazzie, 1991). However, even though the position bias is systematic, it is not the only cue for stress as it only covers less than one third of all cases perceived as stressed by the listeners. This translates to a mean pair-wise agreement rate of $\kappa = 0.20$ (± 0.13) across the listeners if only the last word of each sentence is always hypothesized as stressed. Closer analysis reveals that there are major individual differences between listeners with respect to word position, as can be observed from Fig. 2 showing the listener-specific agreement rates with the "last-word-is-always-stressed model". For four of the listeners, the word position seems to be totally irrelevant while the agreement with listener number three is almost $\kappa = 0.40$.

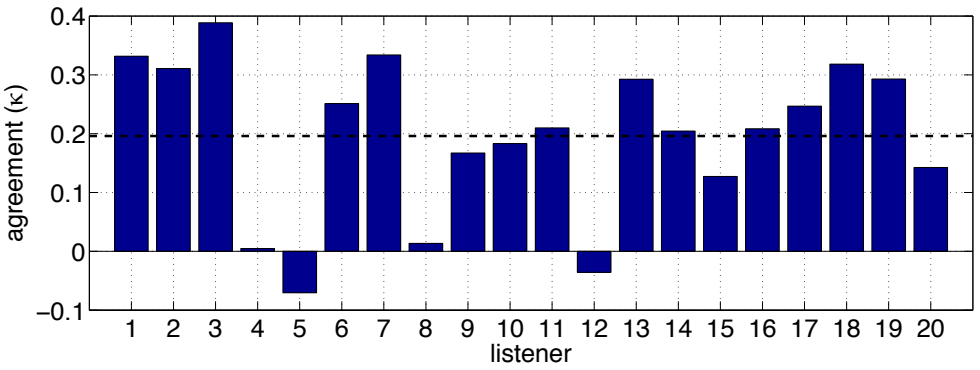

Fig. 2. Annotator-specific agreement rates κ with respect to a model that always chooses the last word of the sentence as stressed. The mean agreement is shown with a horizontal dashed line.

In summary, the average agreement of $\kappa = 0.42$ across all annotators is significantly above chance level and is at the boundary of "fair" and "moderate" agreement according to the Landis and Koch (1977) interpretation of the Fleiss kappa measures. It is also basically the same agreement rate ($\kappa \approx 0.40$) observed in two other studies on prominence perception in American English using native listeners (Mo, Cole, & Lee, 2008; You, 2012). This finding, together with the result that L1 English and Finnish listeners do not differ in their sentence stress judgements, indicates that the current listening test data are representative and provide a reasonable baseline annotation against which the model output can be compared in the simulations carried out in the next sections.

4. Statistical modeling of the prosodic trajectories

The aim of the computational model is to simulate human behavior in the stress perception task by marking words as stressed if the temporal evolution of the prosodic features is unpredictable during the words. To measure the probability of the features, a statistical model operating on a set of acoustic features is needed. In the current work, we study the statistical learning for four main features: F0, signal energy, spectral tilt, and full-band short-term spectrum. In addition, word duration (in seconds) is used as a fifth feature but is not subject to learning since the word boundaries are not known to the algorithm during the training period. Instead, we simply compare the four main features to a duration-based model that favors long words over short in order to understand the additional contribution of the acoustic features. The motivation for the first three features comes from the research on acoustic correlates of prosody. The spectrum is used as a reference feature that contains both segmental and supra-segmental aspects of the signal and it is represented using Mel-frequency cepstral coefficients (MFCCs; Davis & Mermelstein, 1980), a standard spectral feature in automatic speech recognition.

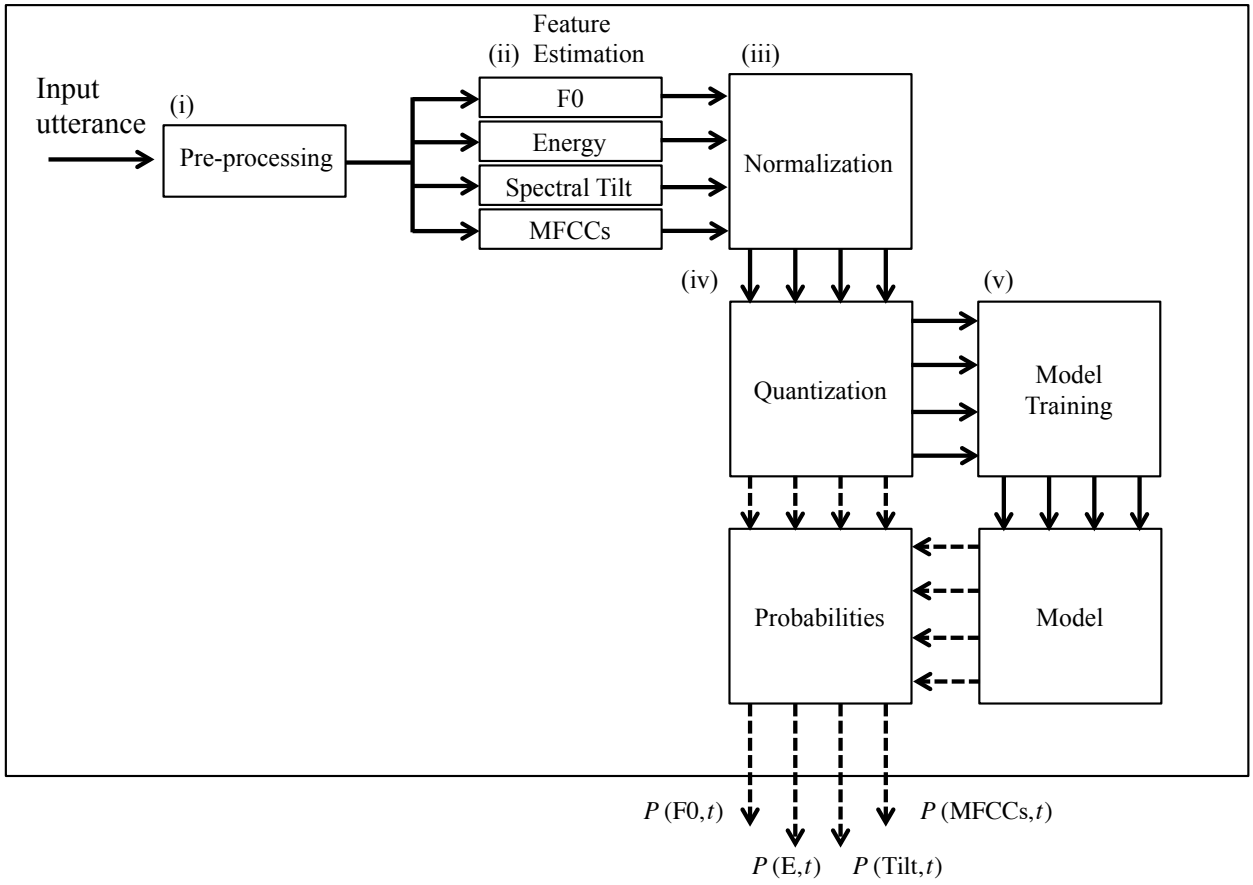

Fig. 3. Overview of the computational model.

An overview of the model is shown in Fig. 3. The model consists of the following main blocks: (i) pre-processing, (ii) feature extraction, (iii) feature normalization, (iv) quantization, and (v) n-gram parameter estimation (during training) or n-gram probability estimation (testing).

4.1. Pre-processing, feature extraction and quantization

As a first pre-processing step, the speech data were downsampled from 44.1 to 8 kHz for F0 and energy computation and to 16 kHz for MFCCs and tilt computation. The three prosodic features and the full spectrum of the speech signals were then computed in 25-ms windows with 10-ms overlap. F0 contours for the voiced segments were extracted from each utterance using the

YAAPT algorithm (Zahorian & Hu, 2008). In order to preserve temporal continuity of the F0 contours, the contours during unvoiced segments were generated by linear interpolation from the neighboring voiced F0 values (see Fig. 4). The energy envelope was computed as

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

where x is the speech waveform at 8 kHz and n_1, n_2 define the beginning and end of the analysis window respectively. The MFCCs were obtained by computing the logarithmic Mel-scale spectrum of a Hamming windowed signal and taking the discrete cosine transform on the resulting spectrum (Davis & Mermelstein, 1980). Spectral tilt was represented by the first MFCC coefficient (see Tsiakoulis, Potamianos, & Dimitriadis, 2010) while the full band spectrum was represented using the 12 first MFCC coefficients.

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

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

In the equation, f denotes the feature value at time t, f' the normalized feature value while max(f) and min(f) refer to the maximum and minimum values of the feature, respectively, during the given utterance (see Imoto et al., 2002). Note that the min-max normalization effectively removes information regarding the absolute values of the features during a sentence, forcing even very flat prosodic trajectories to have clear variation across the sentence length. In our unpredictability framework, however, this is not an issue, as the entire idea is to analyze the

evolution of the prosody on the local sentence level, not to detect stress based on the absolute values of the features.

Finally, in order to allow discrete probability modeling of the data, the extracted features were quantized into 32 discrete states, $f'(t) \rightarrow a_t \in \{1, 2, \ldots, 32\}$, one state occurring every 10 ms. The quantization levels were estimated using the k-means algorithm with an initialization using a set of random samples. The number of levels was manually selected as a compromise between the best possible approximation of the feature contours while using the least number of levels, since too detailed quantization would make statistical learning from finite data problematic (see the next section).

4.2. Statistical modeling

Standard n-grams were chosen to model the discretized features since they are possibly the simplest approach that can account for the temporal evolution of the signal. The analysis was limited to n-gram orders of $n = 2$, 3, and 4, where bi-grams ($n = 2$) correspond to the shortest ordered temporal segments available while the four-grams ($n = 4$) are the longest recurring sequences for which probabilities can be reliably estimated from the given amount of training data.

The probabilities for the n-grams were computed from the relative frequencies of different n-tuples in the training data:

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

where C denotes the frequency counts of the n-tuples in the training data and ψ refers to the feature in question (e.g., ψ = F0). During testing, pre-processing is carried out as in training and the probability $P'(t)$ of the features at time t is computed according to Eq. (4), i.e., by summing the log-probabilities over all features ψ of interest.

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

This formulation assumes that the features are independent of each other, which is generally a reasonable assumption for combining F0 (voice source) with the subglottal and vocal tract properties (energy, tilt, full spectrum). Summing of the non-logarithmic probabilities across features was also initially investigated, but the difference to the multiplication (log-sum) did not lead to any meaningful differences in the results. In order to avoid the logarithm of zero for unseen events in the data, zero probability events were replaced with the value log(0.00001). This corresponds roughly to minus four standard deviations below the mean probability of each utterance (assuming a normal distribution where the majority of the data falls between 1–3 standard deviations, this would represent 0.1% of the overall cases). N-gram smoothing techniques for estimating the missing values were not considered in the model, as the tested n-gram orders (n = 2, 3 and 4) did not result in numerous unseen n-grams during testing, and the aim was to keep the model as simple as possible.

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

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

The temporal boundaries of a word, t_1 and t_2, were extracted from the word-level transcription of the CAREGIVER corpus.

4.3. Stress hypothesis generation

In the following simulations, two conditions were considered when generating the stress hypotheses: (i) the so-called *detection task* where the algorithm selects any number of stressed words for each utterance using a dynamic detection threshold on the word score $S(w)$, and (ii) the *forced-choice task* where the algorithm selects the most stressed word in each utterance.

The detection task corresponds to the listening task where the human subjects were asked to select any number of stressed words out of the words present in each sentence. The stress hypotheses $H(w_{ij})$ for each word j in utterance i were generated by finding the points in time where the word-level scores $S(w_{ij})$ fall below a threshold r_i:

$$H(w_{ij}) = \begin{cases} 1, & S(w_{ij}) < r_i, \\ 0, & S(w_{ij}) \geq r_i, \end{cases} \quad (6)$$

where the threshold is defined as

$$r_i = \mu_i - \sigma_i \lambda. \quad (7)$$

In Eq. (7), μ_i and σ_i are the mean and standard deviation of the word scores $S(w_{ij})$ in the utterance, respectively. λ is a hyperparameter that defines how many standard deviations the given word has to be below the mean score in order to be considered as stressed. In practice, the

simulations were run over a number of λ values in order to find the best fit of the model output to the listening test data.

In the forced-choice task, the algorithm simply chose the word with the lowest $S(w_{ij})$ in each utterance as a stress hypothesis:

$$H(w_{ij}) = \begin{cases} 1, & S(w_{ij}) = \min\{\ S(w_{ij}) \mid j \in [1,2,...,M_i]\ \}, \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

4.4. Duration, feature extreme value, and random baseline systems

Word duration is not explicitly used as a feature in the statistical learning model as the model does not have access to the linguistic content of the signals during training and therefore cannot "learn" typical word durations. However, duration still has an impact on the model output since the instantaneous feature probabilities are integrated over the word duration in the evaluation stage. This means that longer words tend to have lower probabilities even if the probabilities in Eq. (4) are totally random. In order to understand how well durational information can explain the behavioral findings, a simple duration-based model was implemented using the same dynamic thresholding as in Eqs. (6–7) but now operating on word duration instead of word scores:

$$H(w_{ij}) = \begin{cases} 1, & D(w_{ij}) < r_i, \\ 0, & D(w_{ij}) \geq r_i, \end{cases} \tag{9}$$

where $D(w_{ij})$ is the duration of word j in utterance i and r_i is the threshold based on Eq. (7), computed using the mean and standard deviation of the word durations in the same utterance.

Another important baseline is to compare relative to the amplitude-dependent acoustic saliency of different features, since the typical characterizations of stress are related to, e.g., high energy or pitch (Terken, 1991; Sluijter & van Heuven, 1996). In this so-called extreme-value baseline, word-specific maximum (or minimum) values for F0, energy, and spectral tilt were computed and the words were then ranked as stress candidates according to these values. In the detection task, the same number of highest ranking words were chosen as the stress hypotheses for the given utterance as were selected by the predictability model at a given threshold level λ while using the same feature. In the forced-choice task, the word with the highest (or lowest) feature value was chosen as the stress hypothesis. The strength of maximum values and minimum values as a cue to sentence stress were investigated separately.

Finally, in order to investigate chance-level performance in the detection and forced-choice tasks, two different random baselines were computed. In the first random baseline (RB), each word in the test utterances was randomly assigned as either stressed or unstressed with the limitation that the number of stressed words should be equal to the ones hypothesized by the model with a given detection threshold λ. In the durational random baseline (DRB) sampling was performed from a word duration distribution so that the probability of a word being assigned as stressed was linearly proportional to the duration of the word. DRB is therefore somewhat similar to the duration model in Eq. (9) but has random variation in the chosen hypotheses, while Eq. (9) always picks the N longest words depending on the given λ in Eq. (7). Both baselines were computed across 50 iterations at each threshold level.

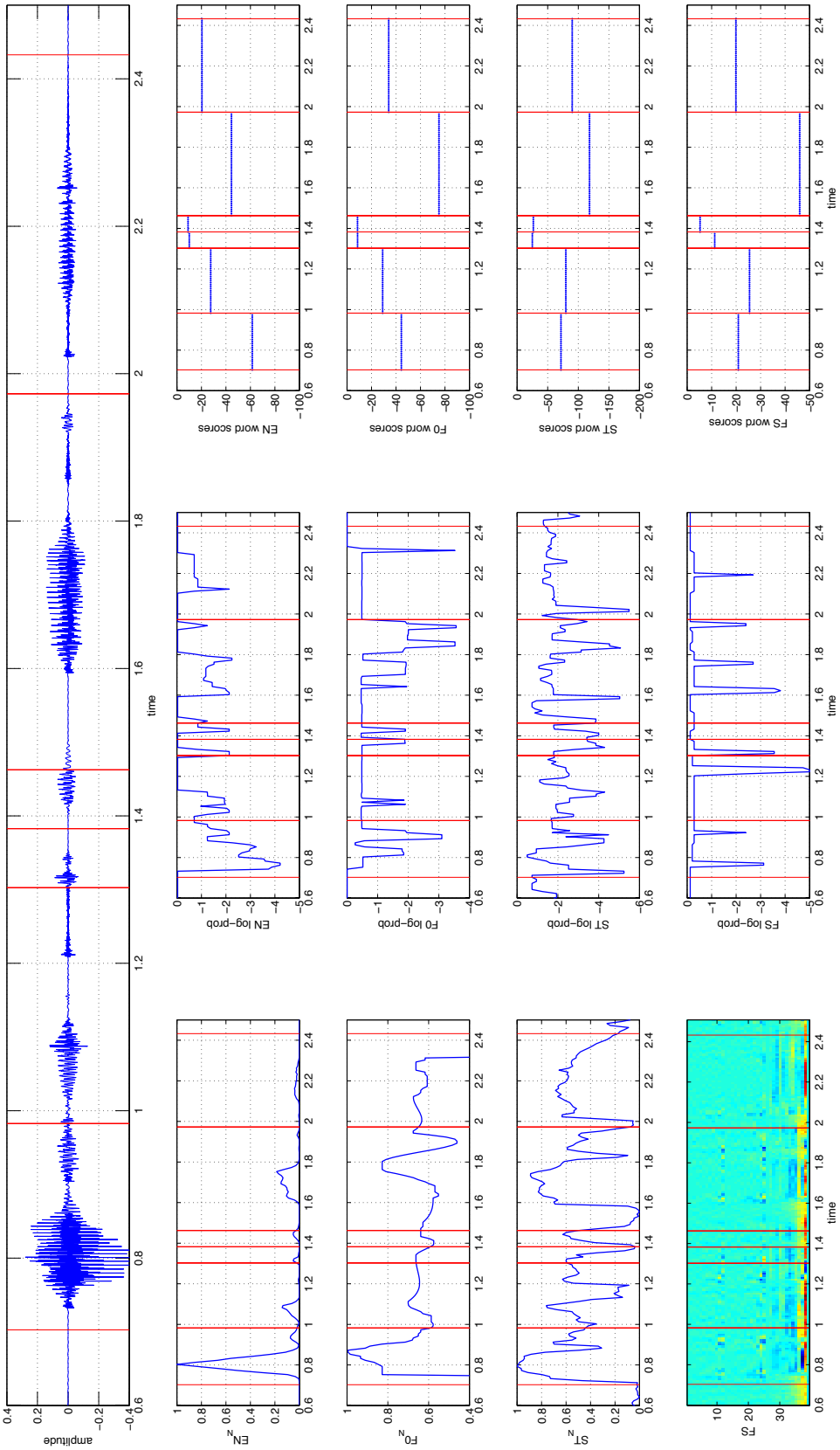
amplitude
time
EN word scores
F0 word scores
ST word scores
FS word scores
EN log−prob
F0 log−prob
ST log−prob
FS log−prob
EN_N
F0_N
ST_N
FS

Fig. 4. Example output of the algorithm for the utterance "*Daddy looks at the **dirty** car*" (from *Speaker 4*), with the word "*dirty*" annotated as stressed by the majority of the annotators. Top row: The original signal waveform. Rows 2-5: The normalized EN_N, $F0_N$, ST_N contours and the full spectrum respectively. Columns 1-3: The features' contours, 2-gram probabilities (with 3-point median filtering for improved visual clarity), and cumulative log-probabilities across the word durations, respectively.

5. Experiments

Two experimental conditions were considered: (i) a detection task, and (ii) a forced-choice task. In both conditions, 4000 training set utterances from the two talkers (see section 3.1.3) were used to train separate n-gram models for the four features (F0, energy, spectral tilt, and MFCCs) and for three different n-gram orders (n = 2, 3, and 4) using speech features quantized into 32 discrete amplitude levels. The models (12 statistical models: $\mathbf{P}_{F0,n}$, $\mathbf{P}_{EN,n}$, $\mathbf{P}_{ST,n}$, $\mathbf{P}_{FS,n}$, $n \in \{2, 3, 4\}$ – see Fig. 4 for an example) and their combinations were then used in order to compute expectations in the test utterances according to two experimental designs explained next.

5.1. Experiment 1: detection task

In the first experiment, the aim was to simulate the listening test scenario where listeners selected zero or more words that they perceived as stressed in the utterance.

After the model was trained, the quantized features from the set of 600 previously unseen human-annotated utterances from the two talkers were used as an input to the model. The probabilities for each feature and all feature combinations were computed based on Eq. (4) using the learned statistical models, yielding a separate probability contour for each feature or

combination for each utterance (see Table 2 for details). Then a binary decision between stressed and non-stressed was performed for each word using the threshold described in Eq. (6). Since the sensitivity of the algorithm depends on the detection threshold λ, the experiment was repeated for all values in the range of $\lambda \in$ [-1.5, 1.5] in steps of 0.05. Word duration was also used as a cue for stress and computed similarly across different λ values (see section 4.4). Fig. 4 shows an example of the different processing stages for the four features.

Table 2. Features and feature combinations used in the experiments.

Feature	Description
EN	energy
F0	fundamental frequency
ST	spectral tilt
FS	full spectrum (MFCCs)
F0+EN	fundamental frequency and energy
F0+ST	fundamental frequency and spectral tilt
ST+EN	spectral tilt and energy
FS+EN	full spectrum and energy
FS+F0	full spectrum and fundamental frequency
FS+ST	full spectrum and spectral tilt
F0+EN+ST	fundamental frequency, energy and spectral tilt
F0+EN+FS	fundamental frequency, energy and full spectrum
FS+EN+ST	full spectrum, energy and spectral tilt
F0+FS+ST	fundamental frequency, full spectrum and spectral tilt

F0+EN+ST+FS	fundamental frequency, energy, spectral tilt and full spectrum
Duration	word duration

The statistical model for all individual features and their combinations was then evaluated for all detection thresholds. Fig. 5 shows the mean pair-wise Fleiss kappa agreement rates of the four individual features: energy, F0, spectral tilt, and full spectrum, while Table 3 shows the results for all individual features and their combinations at different n-gram orders. It is important to note that the model is deterministic given a quantization codebook for the features, making statistical testing between different features difficult. Although the results are averaged across the three n-gram orders in Fig. 5, standard deviations across the n-gram orders are not shown since they are very low for all features ($\sigma \leq 0.01$). Fig. 5 shows that the best correspondence with the human perception of stress is obtained for $\lambda = 0.5$ (see also Table 3). As for the performance of individual features, energy and F0 seem to be the most important cues for stress, leading to agreement levels of $\kappa = 0.45$ and $\kappa = 0.42$ at $\lambda = 0.5$, respectively, being similar to or higher than the mean agreement between the human listeners only ($\kappa = 0.42$). The corresponding values for spectral tilt and the full spectrum are $\kappa = 0.36$ and $\kappa = 0.32$.

In order to confirm that the temporal aspect of the model is relevant to the task, performance of a unigram model ($n = 1$) was also tested for all individual features and their combinations. The general finding was that the performance of the unigram model was worse than the bigram model for all but three cases. Specifically, performance of the best features, namely energy, F0, and their combination were $\kappa = 0.43$, $\kappa = 0.37$, and $\kappa = 0.41$, respectively. Small improvements were observed for spectral tilt ($\kappa = 0.37$), full spectrum ($\kappa = 0.36$), and

their combination ($\kappa = 0.37$). This shows that the predictability in the temporal evolution of the prosodic features contains information that is not available in temporally invariant signal properties, but the unpredictability of the instantaneous values can also provide reasonable estimates of stress in speech.

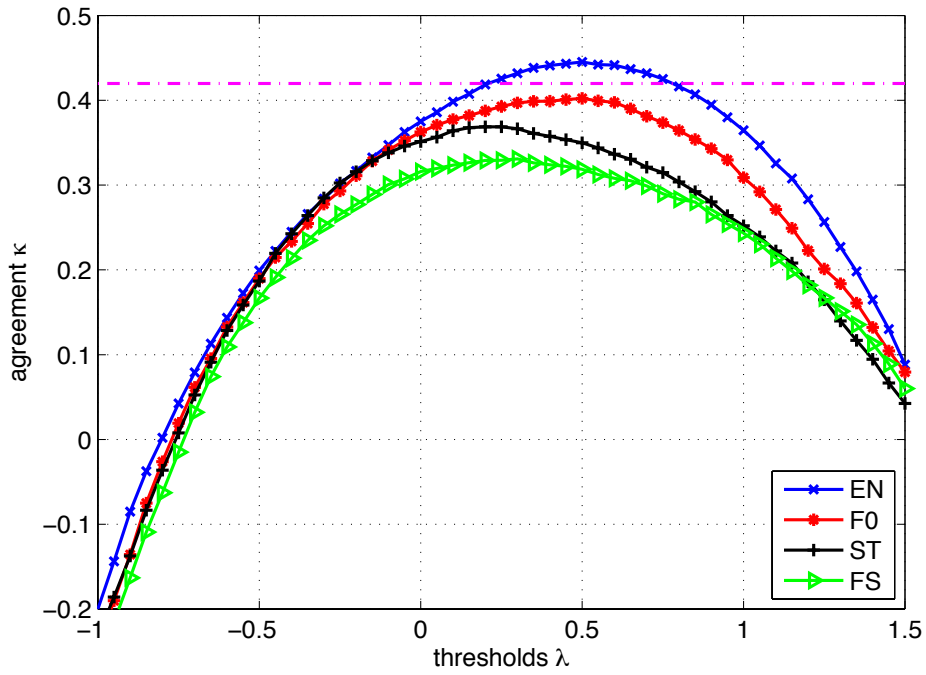

Fig. 5. Mean pair-wise Fleiss kappa between the model's output and the human listeners' annotations as a function of detection threshold λ for the individual features of energy, F0, spectral tilt, and full spectrum, averaged across n-gram orders of $n = 2$, 3, and 4. The horizontal dashed purple line shows the mean pair-wise Fleiss kappa across human listeners.

Table 3. Agreement scores in the detection task for three n-gram orders, all features and feature combinations for a threshold level of $\lambda = 0.5$.

Features	2-gram	3-gram	4-gram

EN	0.45	0.45	0.44
F0	0.41	0.40	0.38
ST	0.37	0.35	0.34
FS	0.32	0.33	0.31
F0+EN	**0.46**	**0.45**	**0.45**
F0+ST	0.39	0.39	0.38
ST+EN	0.43	0.42	0.42
FS+EN	0.43	0.43	0.43
FS+F0	0.38	0.38	0.37
FS+ST	0.35	0.34	0.34
F0+EN+ST	0.43	0.43	0.43
F0+EN+FS	0.44	0.43	0.43
FS+EN+ST	0.41	0.40	0.41
F0+FS+ST	0.38	0.38	0.37
F0+EN+ST+FS	0.42	0.41	0.41
Duration	0.36	0.36	0.36

Word duration is implicitly taken into account in the model through the scoring approach (see section 4.4). Therefore, depending on the feature combination, each computed log-probability score has at least two constituents: (i) the duration and (ii) the feature used for the probability calculation. Fig. 6 shows the performance of the algorithm for the best feature combination of energy and F0 plotted together with the duration-only model and the two random baselines (see also Table 3). The energy and F0 contours were computed by averaging across n =

2, 3, and 4 orders of n-grams similarly to Fig. 5. Word duration alone leads to an agreement level of $\kappa = 0.36$, suggesting that it is a particularly important cue for stress. Addition of the combined effect of energy with F0 leads to an even higher $\kappa = 0.45$. The uniformly sampled random baseline (RB) achieves $\kappa = 0.00$ while, in contrast, the duration random baseline (DRB) reaches a slight agreement of $\kappa = 0.17$. The interplay between word duration and the combined energy and F0 features is further illustrated in Fig. 7 where contributions of duration and the two acoustic features are shown separately for each annotator ($\lambda = 0.5$ and $n = 2$).

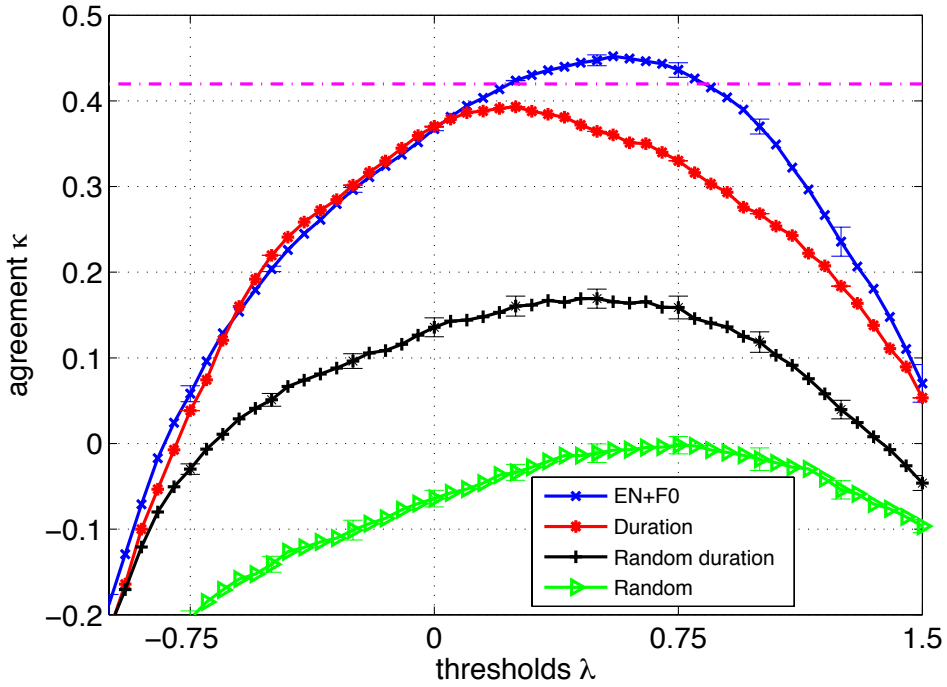

Fig. 6. Mean pair-wise Fleiss kappa between the model's output and the human listeners' annotations as a function of detection threshold λ. The blue line shows the mean model performance for the optimal feature combination of energy and F0 using n-gram orders of $n = 2$, 3, and 4. The red line shows the performance when using only durational information. The black

line shows the durational random baseline (DRB), while the green shows the chance-level performance (RB). Standard deviations are shown with horizontal bars.

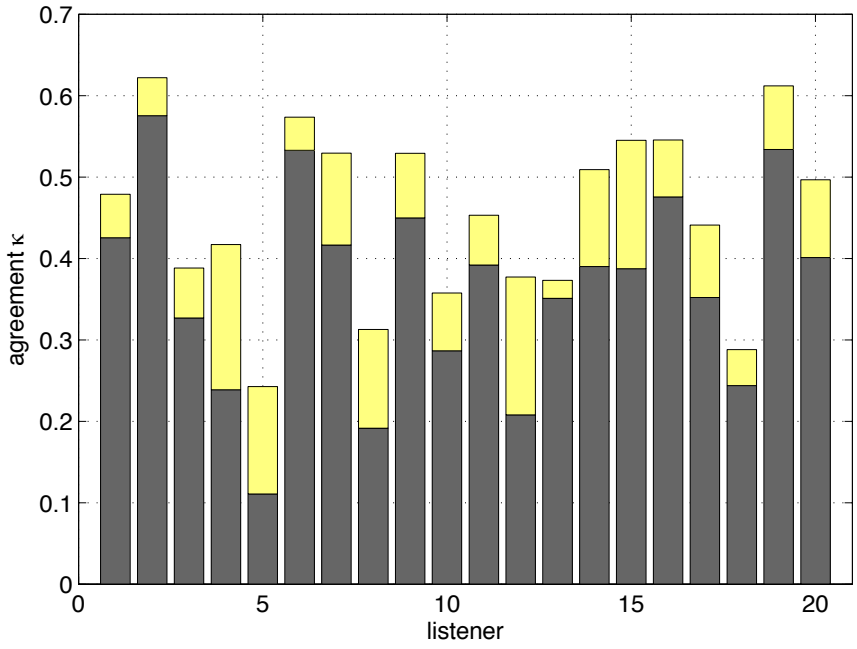

Fig. 7. Pair-wise agreement rates between the model output and each individual annotator for λ = 0.5. The gray bars at the bottom show the contribution of the durational constituent while the yellow bars indicate the contribution of the combined energy with F0.

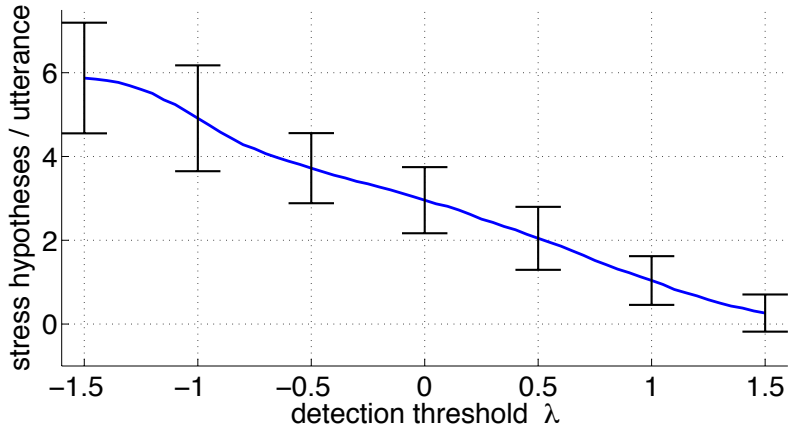

Fig. 8. Average number of hypothesized stress words per utterance as a function of detection threshold λ for the combined F0 and energy features and n-gram order $n = 2$.

Fig. 8 shows the number of generated stress word hypotheses per sentence as a function of the detection threshold λ for the combination of energy and F0. As expected, the number of stress hypotheses increases as the threshold decreases and there are on average 2 (± 0.75) stress hypotheses per utterance at the optimal threshold of $\lambda = 0.5$.

It was also found that if the word scores were normalized by the linear durations of the words, the agreement levels of all features and their combinations were almost random. Since the features still outperform the model using only durational information, this finding suggests that the features interact with the durational information. Another possibility is that the word duration should not be treated as a linear measure, but the normalization should be in terms of, for instance, the logarithmic duration. However, the distribution of probabilities produced by the underlying statistical model and the integration of these probabilities over time results in a complex interaction whose analysis is beyond the scope of the current work. For now, it suffices to say that the predictability of the features, especially energy and F0, contribute significantly to the explanatory power of the model in addition to the duration information alone.

5.1.1. Feature predictability versus feature amplitude

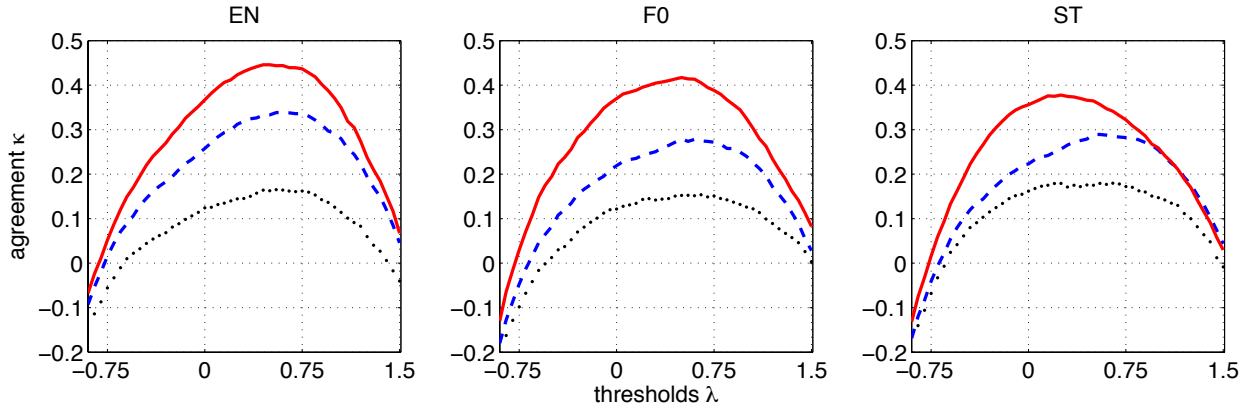

Fig. 9. Comparison of the extreme-value baseline (blue dashed line: max values, black dotted line: min values) to the predictability model (red solid line) using the three main features of energy (left), F0 (middle), and spectral tilt (right). Mean pair-wise Fleiss kappa between the model's output and the human listeners' annotations is shown as a function of the detection threshold λ and using an n-gram order $n = 2$.

A central question with respect to the predictability hypothesis is whether the local extreme feature values such as high pitch or energy also provide equally relevant cues for sentence stress. Therefore the extreme-value baseline was compared to the algorithmic output for $n = 2$ and using the three main features of F0, energy and spectral tilt (Fig. 9). The comparison shows that the feature maxima seem to correlate with the stressed words but the overall performance level is much lower than that achieved by the predictability approach. Specifically, the maximum value baseline for energy reaches $\kappa = 0.33$ (at $\lambda = 0.5$) whereas the corresponding value using the unpredictability algorithm is $\kappa = 0.45$. Similarly, the maxima for F0 gives $\kappa = 0.28$ and spectral tilt $\kappa = 0.29$ as compared to $\kappa = 0.42$ and $\kappa = 0.36$ obtained with the statistical model. This shows that the overall unpredictability of the prosodic trajectories during words is a more robust indicator of sentence stress than the actual values that the prosodic features take during the

sentence. It is important to note that the examined amplitude-based model has been compared against the surprisal-based model only with respect to individual prosodic features, and not in the case where all different features are combined. This is due to the inherent challenges in combining amplitudes of qualitatively different features into a single representative amplitude measure. In contrast, combination in the probability domain is well defined under the assumption of mutually independent features (see Eq. (4)).

5.1.2. The effects of listener habituation and fatigue

The effects of human listener habituation and fatigue to the results were also investigated. One hypothesis was that the listeners might have initially annotated stress based on their intuition and, as the task proceeded, they might have fixated increasingly more on a specific annotation strategy or simply become fatigued by the relatively long task. In order to measure these effects, the temporally local pair-wise agreement rates across listeners and between the model and the listeners were computed in a sliding time window of 75 subsequent utterances. Since the overall agreement rate was found to be much higher for the female talker than the male one, but otherwise showed similar pattern over time, the agreements with respect to both talkers were pooled together. Fig. 8 shows the mean pair-wise agreement rates between listeners and the agreement rates between the listeners and the model as a function of time, measured from the beginning of talker-specific signals during the annotation procedure.

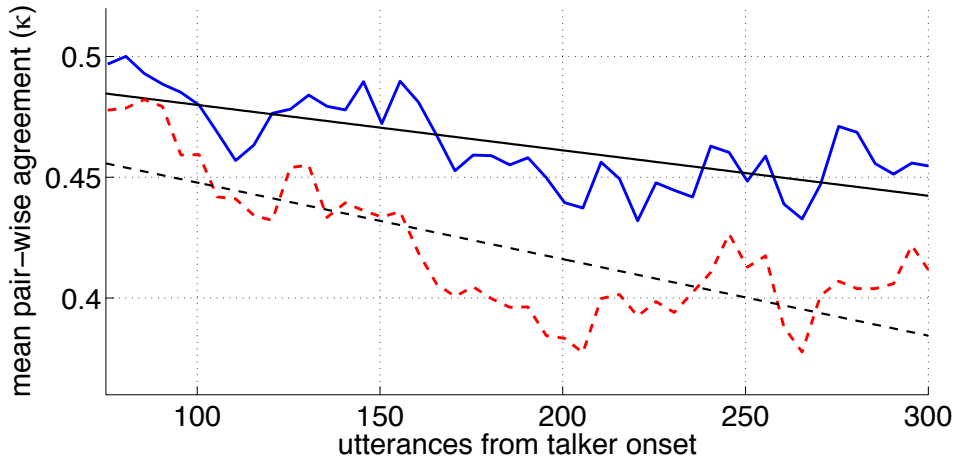

Fig. 10. Mean pair-wise agreement between the model and the listeners (blue solid line) and mean pair-wise agreement across the listeners (red dashed line) as a function of time. The corresponding least-square fits are shown with black lines.

As can be observed, both agreement levels decrease with an increasing number of perceived utterances. Pair-wise agreement between the listeners is significantly higher for the first 75 utterances than the last 75 utterances per talker ($\kappa = 0.48$ versus $\kappa = 0.41$, $p = 0.0001$, one-tailed t-test). Similarly, the agreement between the model and the listeners is higher at the beginning than at the end of the task ($\kappa = 0.50$ versus $\kappa = 0.45$, $p = 0.0271$, one-tailed t-test). As expected, the model agreement also correlates with the inter-annotator agreement rate ($r = 0.88$, $p_r < 0.001$). As the listeners' annotation strategies diverge (or become more noisy) towards the end of the listening task, this also necessarily reduces the agreement between the model and the human listeners. Given the current data, it is impossible to infer the cause for the increasing disagreement. However, the findings show that the model best fits to human perception at an early stage where the listeners are least likely to have established idiosyncratic strategies for the annotation task and should be instead relying on their initial impressions of prominence in the sentences.

5.1.3. Conclusions for experiment 1

Overall, the basic findings from the detection task seem to provide support for the predictability hypothesis. The words that contain the most unpredictable prosodic trajectories within an utterance were shown to correlate best with the human perception of stress in the same utterances. The match is also better for the stress markings made by the listeners in the beginning of the listening test than later when the listeners have more likely adapted some explicit or implicit strategy to perform the task for the given talkers. In contrast, stress hypotheses generated on the basis of highest F0, energy, or tilt during the utterance fall far behind the predictability and duration cues in the task.

5.2. Experiment 2: forced-choice task

The second experiment focused on a scenario where only the most stressed word is selected for each utterance. In order to simulate this situation, the word-level annotations from all listeners were summed resulting in 0-20 stress votes for each word in the data. For each utterance, the word receiving the highest number of votes was then marked as stressed in order to form the reference for the task (two or more words in an utterance never received the same maximum number of votes in the present data).

In this experiment, the task of the algorithm was to simply select the word with the lowest score across each utterance as stressed (see section 4.3). The agreement between the algorithm and the reference was then computed and the results are summarized in Table 4. The best performance is again achieved for the combination of energy with F0, reaching $\kappa = 0.55$ (377 out of 600 stress words detected correctly). The corresponding value for the duration model is $\kappa = 0.37$ (285/600). Note that there is no reference pair-wise agreement level between the human listeners for this experiment as the listeners were not explicitly asked to rank the words according to their relative prominence.

Table 4. Agreement scores in the forced-choice task for three n-gram orders and for all features and feature combinations.

Features	2-gram	3-gram	4-gram
EN	0.52	0.48	0.48
F0	0.42	0.41	0.39
ST	0.33	0.32	0.28

FS	0.30	0.31	0.30
F0+EN	**0.55**	**0.52**	**0.51**
F0+ST	0.35	0.33	0.33
ST+EN	0.45	0.44	0.42
FS+EN	0.47	0.48	0.48
FS+F0	0.36	0.35	0.36
FS+ST	0.31	0.31	0.29
F0+EN+ST	0.46	0.45	0.45
F0+EN+FS	0.48	0.47	0.48
FS+EN+ST	0.43	0.40	0.41
F0+FS+ST	0.33	0.30	0.30
F0+EN+ST+FS	0.41	0.40	0.40
Duration	0.37	0.37	0.37

In order to get an overview of the interplay between duration and the rest of the features, all 15 feature combinations were averaged across the n-gram orders of n = 2, 3, and 4 and then the agreement level of the duration feature alone ($\kappa = 0.37$) was subtracted from all other values. Finally, the resulting values were scaled to have a maximum of one through division with the maximum value, revealing the contribution of all 15 feature combinations evaluated relative to duration (see Fig. 11). Fig. 11 shows that energy and F0 contribute the most to the performance whereas spectral tilt and full spectrum have a negative effect.

In order to understand the effect of extreme feature values in the forced-choice task, the word with the highest feature value in each utterance was selected as a hypothesis in the

extreme-value baseline (see section 4.1.4). It was again observed that high feature values seem to correlate with the perception of stress but with lower agreement levels than the predictability based model. In particular, detection based on energy maxima led to $\kappa = 0.41$ whereas the corresponding value using the unpredictability model was $\kappa = 0.52$. Corresponding values for F0 and spectral tilt were $\kappa = 0.38$ and $\kappa = 0.32$ while the unpredictability model achieved $\kappa = 0.42$ and $\kappa = 0.33$ for the same features, respectively. Finally, the lowest feature values in the utterances were also investigated and shown to produce very low agreement rates for all features similarly to experiment 1. For instance, the agreement using minimum energy detection was $\kappa = 0.17$ while stress detection based on the shortest words achieved $\kappa = -0.20$.

In conclusion, the results of the forced-choice task align with those of the detection task, showing that the prosodic unpredictability seems to be a consistent predictor of sentence stress even when only the most stressed word is considered for each sentence.

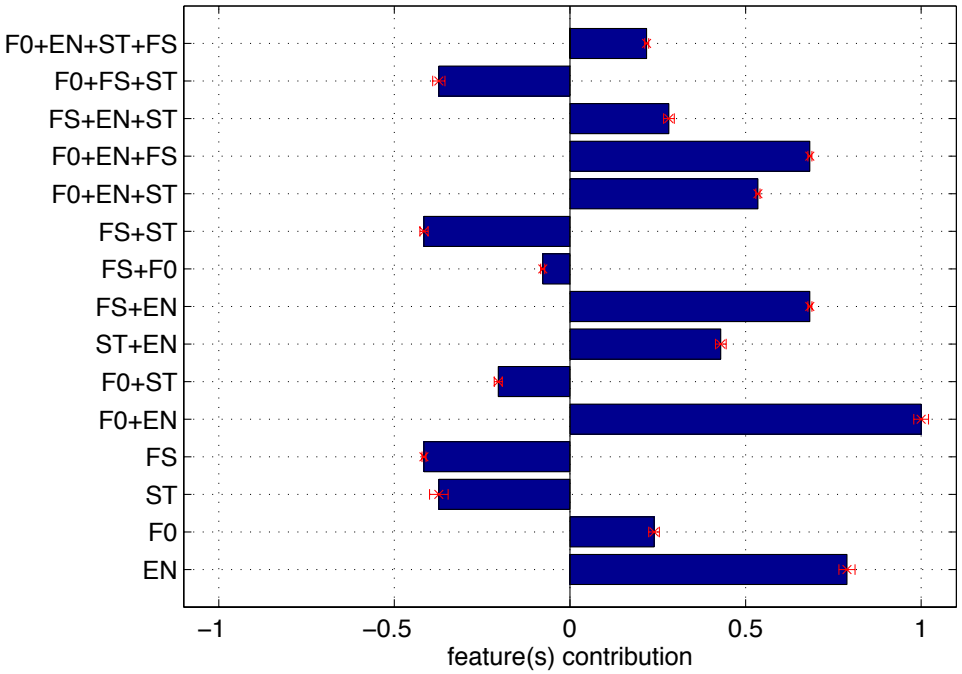

Fig. 11. Normalized and scaled ([-1, 1]) contribution of each individual feature and feature combination in the algorithmic performance evaluated relative to the durational feature and averaged across n-grams orders of n = 2, 3, and 4.

6. Discussion

The present work examined the connection between sentence stress and the temporal unpredictability of prosodic features. Specifically, the work demonstrates that the perception of stress correlates with the word-level unpredictability as measured by the temporal unexpectedness of certain prosodic features. This result provides support to the idea that stimulus unpredictability, auditory attention, and stress perception may be connected and facilitated by statistical learning mechanisms. This converges with the earlier finding of Astheimer and Sanders (2011) who showed that attention is directed to temporal segments of speech that contain unpredictable content. The present computational simulations also show that it is possible to detect stress in an unsupervised manner without having access to a speech corpus with labeled stress markings. Overall, the work provides first evidence that a statistical learning mechanism focused on the regularities and irregularities of the prosodic patterns can be sufficient for stress perception on the sentence level.

6.1. Review of findings

Four acoustic features and word duration were studied with respect to the perception of sentence stress by human listeners. Energy, F0, and duration were found to be the strongest indicators of stress. These results are not surprising and are supported by a large part of the literature (e.g., Lehiste & Peterson, 1959; Fry, 1955; see also Cutler, 2005, for a review). Moreover, combination of the three features seems to best explain human perception than any of these

features alone. According to Lieberman (1960), there are error-correcting trading relationships between these features so that one feature may compensate for another in signaling stress, thereby allowing for more robust communication of stress. From speech production point of view, it may be that the (un)predictability of some features is easier to produce in some articulatory and sentential contexts than in some others.

In contrast, the spectral tilt and spectral envelope were found to be less indicative of sentence stress. Earlier studies on spectral tilt have not reached a consensus on its role since there seem to be contradicting findings on its reliability when applied to different languages. For instance, in their study on lexical stress in Dutch, Sluijter and van Heuven (1996) found spectral tilt to be a reliable cue for stress whereas a later study by Campbell and Beckman (1997) could not replicate the finding for American English (see also Ortega-Llebaria & Prieto, 2010, and references therein). In the present study, the intrinsic inclusion of durational information in all other features makes it difficult to evaluate the contribution of spectral tilt in child-directed British English alone since the overall agreement levels achieved using tilt predictability are close to those of using word duration alone. One possibility for the absence of obvious benefit from tilt may be due to the coupling between the tilt and the segmental content of speech (e.g., flatter spectrum for unvoiced than voiced segments). As the present model disregards the segmental content of the input, tilt cannot be normalized with respect to the underlying phonetic units. Modeling of tilt separately for different underlying segmental units could reveal a more systematic relationship to prominence. However, this would make the present model much more complicated, either requiring an access to a linguistically defined ground truth during learning or by modeling the prosodic features in the context of different concurrent full-band spectral envelopes.

As for the full-band spectrum, the agreement levels are notably lower than those obtained using the duration information alone, suggesting that it is not a very consistent cue for stress. This is an expected result since the spectral envelope mainly carries information regarding the phonetic structure of the language and is therefore inherently dependent on what is said in addition to how it is said. For this reason, the full spectrum has not been studied as a primary correlate of stress but has sometimes been used in supervised algorithms for stress detection (see, e.g., Lai et al., 2006). In our experiments, the full spectrum was simply used as a reference feature in order to see how the lexical and segmental content of the utterances interact with stress perception and the results confirmed that this interaction is weak at best (see Fig. 11).

As for the parameter sensitivity of the model, the model agreement with the human listeners was examined for different n-gram orders. Temporal context sizes of $n = 1, 2, 3$, and 4 (25-55 ms) were used in the experiments, representing the feature trajectories at different lengths and thereby with different details. The best performance for most features and feature combinations was achieved for $n = 2$, with the agreement levels deteriorating slightly with increasing order. The decreasing performance with increasing n-gram orders may be associated with an exponential increase in the number of model parameters, making reliable estimation of n-gram probabilities more difficult from the finite training data (note that the number of possible n-grams is Q^n where Q is the number of discrete amplitude levels).

Regarding the two experimental conditions, both setups examined the relationship between the unexpectedness of the prosodic events in speech and the perception of sentence stress. In the first setup, the best performance ($\kappa = 0.46$), slightly exceeding the average agreement rate among the different annotators ($\kappa = 0.42$), was achieved for a detection threshold

of $\lambda = 0.5$. A closer analysis revealed that the precision of the stress hypotheses increases with an increasing threshold, confirming that the lower word scores seem to be associated with stress. However, with a very high value for λ, the model simply starts to miss many of the stressed words, leading to decreasing overall agreement rate with the human listeners. The second experiment examined how the lowest word score during an entire utterance correlates with the most stressed word in the utterance. The results again showed a high agreement level ($\kappa = 0.55$), further suggesting that there is a strong relationship between words containing the most unexpected prosodic trajectories and that of the perception of stress. Unpredictability is known to be an important cue for attention as it can orient perception towards highly informative events in the environment. Based on our results, sentence stress seems to be guided by a similar mechanism.

Finally, cross-linguistic effects in the perception of stress were investigated by collecting data from both L1 Finnish and L1 UK English listeners. The results did not indicate significant differences in stress perception between the two groups, agreements within a language group being similar to the agreements across language groups. This might be attributed to the high English proficiency of the Finnish listeners (equal to C1&C2 CEF) or to the similarity of the sentence-level stress cues in the two languages. Even though Finnish is a quantity language where segmental durations are used to distinguish word meanings (Järvikivi, Vainio, & Aalto, 2010), it has a systematic trochaic stress pattern that is also prevalent in spoken English. In addition, the sentence-level prosodic patterns of Finnish generally follow similar behavior with those of English. For instance, Suomi, Toivanen, and Ylitalo (2003) and Vainio and Järvikivi (2006) report that stress in Finnish is characterized by the magnitude of change in duration, intensity, and fundamental frequency, that is, in the same features that are significant in

American English (e.g., Batliner et al., 2001). Given the current data, it is impossible to separate the effects of across-language similarities in stress patterns from the notable formal training and informal experience in the use of English by the L1 Finnish listeners.

6.2. Implications for statistical learning research

One of the important motivating factors for the current study was the question of whether stress perception can be learned from exposure to speech and the current results seem to point in that direction. Although the study does not prove that human stress perception is necessarily based on the learned statistical predictability of the prosodic features, it shows that a listener taking the statistical learning strategy will perform similarly to an average human listener in the stress perception task.

Importantly, the current model is very similar to the models attempting to describe early word segmentation and recognition based on the transition probabilities between acoustic or linguistic units, and it is in line with the behavioral findings that human infants and adults are sensitive to statistical regularities across different levels of representation and in different sensory domains (e.g., Baldwin et al., 2008; Fiser & Aslin, 2002; Saffran et al., 1996a; Saffran et al., 1996b; Saffran et al., 1999; Romberg & Saffran, 2010). Earlier studies show that bootstrapping of the lexical learning can be achieved by either hypothesizing a word boundary at the points of low predictability or treating high-probability sequences as words (c.f., e.g., Saffran et al., 1996a; Saffran et al., 1996b; Adriaans, 2011; Swingley, 2005), by treating an entire unfamiliar speech pattern as a potential new lexical item (e.g., Räsänen, 2011; Räsänen & Rasilo, 2012), or by relying on the statistical regularities at the acoustic level with the help of cross-situational constraints from the visual domain (Räsänen, Laine, & Altosaar, 2008). The major

difference to the existing accounts on word segmentation is that the current model operates on the statistical predictability of supra-segmental features instead of linguistically motivated units (phones, syllables) or spectral features that are known to convey segmental and lexical information (see also Dimitrova & Turk, 2012, for a study on syllable lengthening and phrasal prominence). Naturally, the human brain has access to all these features in parallel and therefore statistical learning can take place simultaneously at both the segmental and the supra-segmental level. It is tempting to hypothesize that a single generic learning mechanism explaining perceptual phenomena on linguistic and paralinguistic levels would be more plausible than several different cognitive systems operating on different computational principles. For infant language acquisition, this means that there would be no need for an innate bias or supervised training towards perceiving specific prosodic patterns as stressed, but that the native stress patterns can be learned from linguistic exposure.

6.3. Conclusions and future work

The present study provides first evidence for the hypothesis that unpredictability of the acoustic features in speech conveys sentence stress by capturing the attention of the listener. In future work, extensions to other languages in order to test and validate the hypothesis as a generic sentence stress mechanism would be of particular interest. Another topic of investigation is the amount of exposure required to prime expectations regarding the prosodic features. In the current work, we simply used a fixed-size training material to estimate the parameters of our statistical model whereas the human listeners arrived for the listening test with an extensive pre-existing exposure to stress patterns of English. Moreover, since the listeners encountered the talkers of the speech material for the first time during the annotation process, their gradual adaptation to the talkers or their speaking style might provide a source for the increasing disagreement

between the listeners over time. A better understanding of the time scales and required language exposure needed to develop prosodic expectations for new talkers is therefore an important topic for future research. Also, the experiments should be replicated using adult-directed speech in order to verify generality of the present findings.

One of the predictions of the present work is that unpredictability in any domain (e.g., lexical or grammatical level) should cause similar attentional orientation behavior to what is observed with the prosodic features as long as the change in the signal predictability is of similar magnitude at both levels. However, since there is typically more uncertainty involved at the lexical level than at the supra-segmental level (e.g., the next words spoken by a talker versus their expected pitch or energy), violations to the expectations at the supra-segmental level may be more distinct and therefore have a stronger effect on the perceptual orientation. However, it should be possible to test this hypothesis in controlled listening scenarios where the predictability at both lexical and supra-segmental level can be equated and then systematically manipulated (see also Aylett & Turk, 2004, for a similar idea). Such a study could also shed some light on the open question of whether the strength of stress perception is dependent on the degree of unpredictability in a graded manner (i.e., the higher the unpredictability the stronger the perception of stress), or whether auditory attention is better characterized in terms of binary stress/no-stress decisions.

Acknowledgements

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

acquisition". We would also like to thank Ellen Gurman Bard and the three anonymous reviewers for their invaluable comments and all who took part in the listening tests.

References

Adriaans, F. (2011). *The induction of phonotactics for speech segmentation. Converging evidence from computational and human learners* (Doctoral Dissertation). Retrieved from Utrecht University Repository.

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 7th International Conference on Language Resources and Evaluation (LREC-2010)* pp. 1062–1068. Malta.

Astheimer, L. B., & Sanders, L. D. (2009). Listeners modulate temporally selective attention during natural language processing. *Biological Psychology, 80*, 23–34. doi:10.1016/j.biopsycho.2008.01.015.

Astheimer, L. B., & Sanders, L. D. (2011). Predictability affects early perceptual processing of word onsets in continuous speech. *Neuropsychologia*, *49*, 3512–3516. doi:10.1016/j.neuropsychologia.2011.08.014.

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. doi:10.1177/00238309040470010201.

Baldwin, D., Andersson, A., Saffran, J., & Meyer, M. (2008). Segmenting dynamic human action via statistical structure. *Cognition, 106*, 1382–1407. doi:10.1016/j.cognition.2007.07.005.

Batliner, A., Buckow, J., Huber, R., Warnke, V., Nöth, E., & Niemann, H. (2001). Boiling down prosody for the classification of boundaries and accents in German and English. *Proceedings of the 2nd Annual Conference of the International Speech Communication Association (Interspeech-2001)*, pp. 2781–2784, Aalborg, Denmark.

Birch, S., & Clifton, C. (1995). Focus, accent, and argument structure: Effects on language comprehension. *Language and Speech*, *38*, 365–391. doi: 10.1177/002383099503800403.

Blakemore, C., & Cooper, G. (1970). Development of the brain depends on the visual environment. *Nature,* 228, 477–478. doi:10.1038/228477a0.

Bock, J. K., & Mazzella, J. R. (1983). Intonational marking of given and new information: Some consequences for comprehension. *Memory & Cognition*, *11*, 64–76. doi:10.3758/BF03197663.

Bolinger, D. L. (1964). Intonation: Around the edge of language. *Harvard Educational Review*, *34*, 282–296.

Calhoun, S. (2007). Predicting focus through prominence structure. *Proceedings of the 8th Annual Conference of the International Speech Communication Association (Interspeech-2007)*, pp. 622–625, Antwerp, Belgium.

Calhoun, S. (2010). The centrality of metrical structure in signaling information structure: A probabilistic perspective. *Language*, *86*, 1–42. doi: 10.1353/lan.0.0197.

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, Stockholm, Sweden.

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).

Chaolei, L., Jia, L., & Shanhong, X. (2007). English sentence stress detection system based on HMM framework. *Applied Mathematics and Computation*, *185*, 759–768. doi:10.1016/j.amc.2006.06.081.

Chrabaszcz, A., Winn, M., Lin, C. Y., & Idsardi, W. J. (2014). Acoustic Cues to Perception of Word Stress by English, Mandarin and Russian Speakers. *Journal of Speech, Language, and Hearing Research*, 57, 1468–1479. doi:10.1044/2014_JSLHR-L-13-0279.

Corbetta, M., & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature reviews neuroscience*, *3*, 201–215. doi:10.1038/nrn755.

Cutler, A. (2005). Lexical Stress. In D. B. Pisoni & R. E. Remez (Eds.), *The Handbook of Speech Perception* (pp. 264–289). Malden, MA, Oxford, and Carlton, Victoria: Blackwell publishing.

Cutler, A., & Darwin, C. J. (1981). Phoneme-monitoring reaction time and preceding prosody: effects of stop closure duration and of fundamental frequency. *Perception and Psychophysics*, *29*, 217–224. doi:10.3758/BF03207288.

Cutler, A., & Foss, D. J. (1977). On the role of sentence stress in sentence processing. *Language and Speech*, *20*, 1–10. doi:10.1177/002383097702000101.

Cutler, A., Oahan, D., & Donselaar, W. (1997). Prosody in the comprehension of spoken langauge: A literature review. *Language and Speech*, *40*, 141–201. doi:10.1177/002383099704000203.

Davis, S. B., & Mermelstein, P. (1980). Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *28*, 357–366. doi:10.1109/TASSP.1980.1163420.

Dimitrova, S., & Turk, A. (2012). Patterns of accentual lengthening in English four-syllable words. *Journal of Phonetics*, *40*, 403–418. doi:10.1016/j.wocn.2012.02.008.

Endress, A. D., & Hauser, M. D. (2010). Word segmentation with universal prosodic cues. *Cognitive Psychology*, *61*, 177–199. doi:10.1016/j.cogpsych.2010.05.001.

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 the 14th Annual Conference of the International Speech Communication Association (Interspeech-2013)*, pp. 778–782, Lyon, France.

Feldman, N. H., Griffiths, T., & Morgan, J. L. (2009). Learning phonetic categories by learning a lexicon. *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, pp. 2208–2213, Amsterdam, Netherlands.

Fernald, A., & Mazzie, C. (1991). Prosody and focus in speech to infants and adults. *Developmental Psychology*, *27*, 209–221. doi:10.1037/0012-1649.27.2.209.

Fiser, J., & Aslin, R. N. (2002). Statistical learning of higher-order temporal structure from visual shape sequences. *Journal of Experimental Psychology*, *28*, 458–467. doi:10.1037/0278-7393.28.3.458.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*, 378–382. doi:10.1037/h0031619.

Folk, C. L., Remington, R. W., & Johnston, J. C. (1992). Involuntary covert orienting is contingent on attentional control settings. *Journal of Experimental Psychology: Human Perception and Performance, 18*, 1030–1044. doi:10.1037/0096-1523.18.4.1030.

Fry, D. B. (1955). Duration and intensity as physical correlates of linguistic stress. *Journal of the Acoustical Society of America*, *27*, 765–768. doi:10.1121/1.1908022.

Gussenhoven, C. (2011). Sentential prominence in English. In M. van Oostendorp, C. J. Ewen, E. Hume, & K. Rice (Eds.), *The Blackwell Companion to Phonology* (pp. 2780–2806). Malden, MA and Oxford: Wiley-Blackwell.

Hirschberg, J. (1993). Pitch accent in context predicting intonational prominence from text. *Artificial Intelligence*, *63*, 305–340. doi:10.1016/0004-3702(93)90020-C.

Imoto, K., Dantsuji, M., & Kawahara, T. (2000). Modelling of the perception of English sentence stress for computer-assisted language learning. *Proceedings of the 1st Annual Conference of the International Speech Communication Association (Interspeech-2001)*, pp. 175–178, Beijing, China.

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 3rd Annual Conference of the International Speech Communication Association (Interspeech-2002)*, pp. 749–752, Denver, CO.

Itti, L., & Baldi, P. (2005). Bayesian surprise attracts human attention. In Y. Weiss, B. Schölkopf, & J. Platt (Eds.): *Proceedings of the 19th Annual Conference on Advances in Neural Information Processing Systems (NIPS-2005)* (pp. 547–554). Cambridge, MA: MIT Press.

Itti, L., & Baldi, P. (2009). Bayesian surprise attracts human attention. *Vision Research*, *49*, 1295–1306. doi:10.1016/j.visres.2008.09.007.

Järvikivi, J., Vainio, M., & Aalto, D. (2010). Real-time correlates of phonological quantity reveal unity of tonal and non-tonal languages. *PloS one*, *5*, e12603. doi:10.1371/journal.pone.0012603.

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*, 1009–1024. doi:10.1109/TASL.2009.2014795.

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. doi:10.1121/1.1923349.

Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., & Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science*, *255*, 606–608. doi: 10.1126/science.1736364.

Laakso, A., & Calvo, P. (2011). How many mechanisms are needed to analyze speech? A connectionist simulation of structural rule learning in artificial language acquisition. *Cognitive Science*, *35*, 1243–1281. doi: 10.1111/j.1551-6709.2011.01191.x.

Ladd, D. R. (2008). *Intonational phonology*. Cambridge: Cambridge University Press.

Lai, M., Chen, Y., Chu, M., Zhao, Y., & Hu, F. (2006). A hierarchical approach to automatic stress detection in English sentences. *Proceedings of the 31st International Conference on Acoustics, Speech and Signal Processing (ICASSP-2006)*, pp. 753–756, Toulouse, France.

Lake, B. M., Vallabha, G. K., & McClelland, J. L. (2009). Modeling unsupervised perceptual category learning. *IEEE Transactions on Autonomous Mental Development*, *1*, 35–43. doi:10.1109/TAMD.2009.2021703.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159–174. doi:10.2307/2529310.

Lehiste, I. (1970). *Suprasegmentals*. Cambridge, Massachusetts: MIT Press.

Lehiste, I., & Peterson, G. E. (1959). Vowel amplitude and phonemic stress in American English. *Journal of the Acoustical Society of America*, *31*, 428–435. doi:10.1121/1.1907729.

Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: A quick and valid Lexical Test for Advanced Learners of English. *Behavior Research Methods, 44,* 325–343. doi:10.3758/s13428-011-0146-0.

Li, K., Zhang, S., Li, M., Lo, W. K., & Meng, H. (2011). Prominence model for prosodic features in automatic lexical stress and pitch accent detection. *Proceedings of the 12th Annual Conference of the International Speech Communication Association (Interspeech-2011),* pp. 2009–2012, Makuhari, Japan.

Lieberman, P. (1960). Some acoustic correlates of word stress in American English. *The Journal of the Acoustical Society of America*, *32*, 451–454. doi:10.1121/1.1908095.

Malisz, Z., & Wagner, P. (2012). Acoustic-phonetic realization of Polish syllable prominence: a corpus study. In D. Gibbon, D. Hirst, & N. Campbell (Eds.), *Speech and Language Technology: Vol. 14/15. Rhythm, melody and harmony in speech. Studies in honour of Wiktor Jassem* (pp. 105–114).

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition, 82*, B101–B111. doi:10.1016/S0010-0277(01)00157-3.

Mehrabani, M., Mishra, T., & Conkie, A. (2013). Unsupervised prominence prediction for speech synthesis. *Proceedings of the 14th Annual Conference of the International Speech Communication Association (Interspeech-2013)*, pp. 1559–1563, Lyon, France.

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 7th International Conference on Spoken Language Processing (ICSLP-2002)*, pp. 745–748, Denver, CO.

Mishra, T., Sridhar, V. K. R., & Conkie, A. (2012). Word Prominence Detection using Robust yet Simple Prosodic Features. *Proceedings of the 13th Annual Conference of the International Speech Communication Association (Interspeech-2012)*, pp. 1864–1867, Portland, Oregon.

Mo, Y., Cole, J., & Lee, E. K. (2008). Naïve listeners' prominence and boundary perception. *Proceedings of the 4th Conference on Speech Prosody*, pp. 735–738, Campinas, Brazil.

Moubayed, S. A., Ananthakrishnan, G., & Enflo, L. (2010). Automatic prominence classification in Swedish. *Proceedings of the 5th International Conference on Speech Prosody*, pp. 1–10, Chicago, IL.

Ortega-Llebaria, M. (2006). Phonetic cues to stress and accent in Spanish. In. M. Diaz Campos (Ed.): *Selected Proceedings of the 2nd Conference on Laboratory Approaches to Spanish Phonetics and Phonology* (pp. 104–118). Cascadilla Proceedings Project, Somerville, MA.

Ortega-Llebaria, M., & Prieto, P. (2010). Acoustic correlates of stress in Central Catalan and Castilian Spanish. *Language and Speech*, *54*, 1–25. doi:10.1177/0023830910388014.

Pannekamp, A., Toepel, U., Alter, K., Hahne, A., & Friederici, A. D. (2005). Prosody-driven sentence processing: An Event-related Brain Potential Study. *Journal of Cognitive Neuroscience*, *17* (3), 1–15. doi:10.1162/0898929053279450.

Ranganath, C., & Rainer, G. (2003). Neural mechanisms for detecting and remembering novel events. *Nature Reviews Neuroscience*, *4*, 193–202. doi:10.1038/nrn1052.

Romberg, A. R., & Saffran, J. R. (2010). Statistical learning and language acquisition. *Wiley Interdisciplinary Review of Cognitive Science, 1*, 906–914. doi:10.1002/wcs.78.

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 (NAACL HLT-2009)*, pp. 81–84, Boulder, CO.

Räsänen, O. (2011). A computational model of word segmentation from continuous speech using transitional probabilities of atomic acoustic events. *Cognition*, *120*, 149–176. doi:10.1016/j.cognition.2011.04.001.

Räsänen, O. (2012). Computational modeling of phonetic and lexical learning in early language acquisition: existing models and future directions. *Speech Communication*, *54*, 975–997. doi:10.1016/j.specom.2012.05.001.

Räsänen, O., & Rasilo, H. (2012). Acoustic analysis supports the existence of a single distributional learning mechanism in structural rule learning from an artificial language. *Proceedings of the 34th Annual Conference of the Cognitive Science Society*, pp. 887–892, Sapporo, Japan.

Räsänen, O., Laine, U. K., & Altosaar, T. (2008). Computational language acquisition by statistical bottom-up processing. *Proceedings of the 9th Annual Conference of the International Speech Communication Association (Interspeech-2008)*, pp. 1980–1983, Brisbane, Australia.

Saffran, J. R., Johnson, E. K., Aslin, R. N., & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition*, *70*, 27–52. doi:10.1016/S0010-0277(98)00075-4.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996a). Statistical learning by 8-month-old infants. *Science*, *274*, 1926–1928. doi:10.1126/science.274.5294.1926.

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996b). Word segmentation: The role of distributional cues. *Journal of Memory and Language*, *35*, 606–621. doi:10.1006/jmla.1996.0032.

Saffran, J. R. (2001). The use of predictive dependencies in language learning. *Journal of Memory and Language, 44*, 493–515. doi:10.1006/jmla.2000.2759.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*, 379–423. doi:10.1145/584091.584093.

Shattuck-Hufnagel, S., & Turk, A. E. (1996). A prosody tutorial for investigators of auditory sentence processing. *Journal of Psycholinguistic Research*, *25*, 193–247. doi:10.1007/BF01708572.

Shields, J. L., McHugh, A., & Martin, J. G. (1974). Reaction time to phoneme targets as a function of rhythmic cues in continuous speech. *Journal of Experimental Psychology*. *102*, 250–255. doi:10.1037/h0035855.

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. doi:10.1121/1.417955.

Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition, 106*, 1558–1568. doi:10.1016/j.cognition.2007.06.010.

Sokolov, E. N. (1963). Higher nervous functions: the orienting reflex. *Annual Review of Psychology, 25*, 545–580. doi:10.1146/annurev.ph.25.030163.002553.

Sridhar, V. K. R., Nenkova, A., Narayanan, S., & Jurafsky, D. (2008). Detecting prominence in conversational speech: pitch accent, givenness and focus. *Proceedings of the 4th Conference on Speech Prosody,* pp. 456–459, Campinas, Brazil.

Steinhauer, K. (2003). Electrophysiological correlates of prosody and punctuation. *Brain and Language*, *86* (1), 142–164. doi:10.1016/S0093-934X(02)00542-4.

Suomi, K., Toivanen, J., & Ylitalo, R. (2003). Durational and tonal correlates of accent in Finnish. *Journal of Phonetics*, *31*, 113–138. doi:10.1016/S0095-4470(02)00074-8.

Sur, M., Garraghty, P. E., & Roe, A. W. (1988). Experimentally induced visual projections into auditory thalamus and cortex. *Science, 242*, 1437–1441. doi:10.1126/science.2462279.

Sussman, E., Ritter, W., & Vaughan, H. G. (1998). Attention affects the organization of auditory input associated with the mismatch negativity system. *Brain research*, *789*, 130–138. doi:10.1016/S0006-8993(97)01443-1.

Swingley, D. (2005). Statistical clustering and the contents of the infant vocabulary. *Cognitive Psychology*, *50*, 86–132. doi:10.1016/j.cogpsych.2004.06.001.

Tamburini, F. (2003). Automatic prosodic prominence detection in speech using acoustic features: an unsupervised system. *Proceedings of the 4th Annual Conference of the*

International Speech Communication Association (Interspeech-2003), pp. 129–132, Geneva, Switzerland.

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. doi:10.1007/s10772-005-4760-z.

Tamburini, F., & Wagner, P. (2007). On automatic prominence detection for German. *Proceedings of the 8th Annual Conference of the International Speech Communication Association (Interspeech-2007)*, pp. 1809–1812, Antwerp, Belgium.

Teinonen, T., Aslin, R. N., Alku, P., & Csibra, G. (2008). Visual speech contributes to phonetic learning in 6-month-old infants. *Cognition, 108*, 850–855. doi:10.1016/j.cognition.2008.05.009.

Terken, J. (1991). Fundamental frequency and perceived prominence of accented syllables. *Journal of the Acoustical Society of America*, *89*, 1768–1776. doi:10.1121/1.401019.

Tsiakoulis, P., Potamianos, A., & Dimitriadis, D. (2010). Spectral moment features augmented by low order cepstral coefficients for robust ASR. *IEEE Signal Processing Letters*, *17*, 551–554. doi:10.1109/LSP.2010.2046349.

Vainio, M., & Järvikivi, J. (2006). Tonal features, intensity, and word order in the perception of prominence. *Journal of Phonetics*, *34*, 319–342. doi:10.1016/j.wocn.2005.06.004.

Wang, D., & Narayanan, S. (2007). An acoustic measure for word prominence in spontaneous speech. *IEEE Transactions on Audio, Speech, and Language Processing, 15*, 690–701. doi:10.1109/TASL.2006.881703.

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

You, H. J. (2012). Determining prominence and prosodic boundaries in Korean by non-expert rapid prosody transcription. *Proceedings of the 6th International Conference on Speech Prosody*, Shanghai, China.

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. doi:10.1121/1.2916590.