Accepted manuscript to be published in *Speech Communication*:
doi:10.1016/j.specom.2016.06.004

3PRO – An Unsupervised Method for the Automatic Detection of Sentence Prominence in Speech

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

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

PO Box 00076, Aalto, Finland. E-mail: firstname.lastname@aalto.fi, phone[*]: +358-50-4080134

Abstract

Automatic detection of prominence in speech has attracted interest in recent years due to its multiple uses in spoken language applications. However, typical approaches require manual labeling of the data that is an expensive and time consuming process, also making the systems potentially specific to the language or speaking style in question. In this paper, we propose a novel unsupervised algorithm for the automatic detection of sentence prominence named 3PRO (Prominence from Prosodic Probabilities; "*three-pro*") that is based on recent findings on human perception of prominence in speech. By combining syllable duration information to the level of surprisal observed in the acoustic prosodic features, the method is capable of estimating prominent words from continuous speech without labeled training data. The algorithm is evaluated by comparing model output to manually transcribed prominence labels on a Dutch and French speech corpus, showing performance levels close to supervised prominence classifiers operating on the same data.

Keywords: Prominence detection; Prominence perception; Speech processing; Speech analysis; Stimulus predictability; Attention

1. Introduction

Speech contains information that extends the content of a written message and includes characteristics such as the identity of the speaker, the emotional state, information status, and intonational patterns. Prosody is the defining organizational property of these characteristics and sentence prominence is one type of prosodic event that is most commonly used in order to refer to the perceptual salience of one or multiple words within a sentence (see, e.g., Cutler, 2005; Kohler, 2008; Wagner et al., 2015). For instance, in natural conversation, it is common for speakers to make some words more prominent in order to highlight information and draw the listeners' attention to specific parts in an utterance. In general, speakers use prosody and prosodic phenomena in order to direct and control aspects of the listeners' perception (Cutler, 1987). Therefore, methods of detecting prominence have potential use in spoken language systems such as in automatic speech recognition (ASR) and speech synthesis, not only to account for natural acoustic variability in the signal, but also to ensure that the intended meaning of the message is properly represented.

In the present paper, we describe a new method for the automatic detection of sentence prominence in speech named 3PRO (Prominence from Prosodic Probabilities; "*three-pro*"). The method is based on unsupervised estimation of the statistical properties of prosodic features combined with duration estimation from syllabic rhythm, providing a labeling-, and potentially, language-independent approach for prominence detection. The method is also applicable with different temporal constraints in the prominence detection task, allowing either purely unsupervised prominence detection, or detection in terms of externally provided temporal units (e.g., word boundaries from ASR). We first review the concepts of prominence and existing approaches for its detection from speech, also describing why stimulus predictability can be used as a basis for prominence detection. This is followed by a description of the proposed 3PRO-algorithm and a set of computational experiments where the method is validated on a Dutch and French speech corpus.

1.1. Background

The role of prosody in speech production is to assist in the communicative function (Cutler, 1987) where a speaker may choose to employ certain prosodic coding in order to serve a particular, linguistic or paralinguistic, intent (Werner & Keller, 1994). Specifically, prosody refers to information that extends the segmental content of utterances (individual phonemes) and spans through longer structures such as those of syllables, words and phrases—therefore commonly referred to as suprasegmental information (Lehiste, 1970). As there are currently several definitions for prosody (see Shattuck-Hufnagel & Turk, 1996), here we use the term with reference to the phonetic characterization where prosody is defined as variation of the acoustic parameters in speech and is believed to signal constituent boundaries and prominence (Shattuck-Hufnagel & Turk, 1996).

Typical acoustic prosodic parameters are the fundamental frequency (F0), energy, duration, and spectral tilt. For prominence, several studies support the role of F0, energy and duration (see, e.g., Lieberman, 1960; Fry, 1955, 1958; Kochanski et al., 2005; Terken, 1991). Moreover, spectral tilt has been shown to be a good correlate of prominence in Dutch (Sluijter & van Heuven, 1996) with, however, fewer studies supporting its role across languages (see, e.g., Campbell, 1995; Campbell & Beckman, 1997, for studies in American English; see also Ortega-Llebaria & Prieto, 2010, for a discussion). As for its function, it has been well established that one of the primary functions of prominence in language production is to convey the information status of words (see, e.g., Calhoun, 2007, 2010a, 2010b). Speakers' intent can therefore be directly reflected on the information status of each word through prosodic coding. On the listeners' side, decoding of prominence-related information may translate into a change of the perceptual orientation or focus. This has important implications for language comprehension as it allows, for instance, rapid and efficient recognition of the word (see, e.g., Cutler, Dahan, & Van Donselaar, 1997) and faster sentence comprehension (see, e.g., Bock & Mazzella, 1983).

Both acoustic (bottom-up) and structural components of the language (top-down) can be used to convey prominence (see, e.g., Wagner, Tamburini, & Windmann, 2012; see also Arnold & Wagner, 2008). Acoustic components refer to the articulatory changes that are possible without altering the literal meaning of the words. In contrast, structural cues to prominence correspond, for instance, to changes in the ordering of individual lexemes in free word order languages (Ladd, 2008). The work of Luchkina and Cole (2014) with Russian suggests that word ordering is an optional resource for encoding prominence, and both acoustic and structural means can be utilized together in order to convey the desired information status of words. One particularly interesting direction in the research of sentence prominence is the investigation of cross-language differences in production and perception. Based on the findings thus far, it seems that languages share the same pool of acoustic-prosodic features for signaling prominence, where, however, the language specific realization may vary (see Koreman et al., 2009, for a study in Norwegian and German). This means that different languages may employ a combination of all or a subset of these features and also control the extent each feature is used in order to convey prominence (see also Andreeva, Barry, & Koreman, 2014, for a multi-language comparison; Endress & Hauser, 2010). These differences may be attributed to the underlying distinctions coming from the dissimilar phonological structures of the languages, and also suggest that prominence perception may depend on experience with the given language. For instance, languages may have different rhythm types (such as stress-timed, syllable-timed, and mora-timed) or use tones in order to distinguish words. Therefore, although the phonological structure of languages varies, the phonetic basis descriptive of prominence seems to be shared. Moreover, the perceptual outcome across all languages seem to be the same, that is, prominence results in the perceptual orientation of the listener to specific parts in the utterance (attentional shift). This suggests that there might be a cognitive mechanism that makes prominence perception possible across languages by combining the general acoustic features of prominence to some type of learning mechanism that is responsible for capturing the language-specific prosodic patterns.

A number of studies have examined the plausibility of cross-language prominence detection. For instance, the study of Moniz et al. (2014) showed that it was possible to train a prominence detection model in American English using the AuToBI prosodic event detection system (Rosenberg, 2010) and apply the same model in a prominence detection task in European Portuguese with performance similar to state-of-the-art results (see, Moniz, Mata, Hirschberg, Batista, Rosenberg, & Trancoso, 2014; see also Rosenberg, Cooper, Levitan, & Hirschberg, 2012; Maier et al., 2009). Fewer studies have investigated the perceptual and cognitive mechanisms responsible for prominence perception. These studies have utilized cognitive-inspired approaches on the basis of modeling attention as the mechanism enabling the perceptual shift, as prominence perception has been associated with the function of attention (see, e.g., Cole, Mo, & Hasegawa-Johnson, 2010; Kalinli & Narayanan, 2009; Kakouros & Räsänen, 2014a). For instance, Cole et al. (2010) concluded that attention and prominence might share a common basis where a word may attract the listener's attention, either as a response to acoustic modulation (signal-based acoustic salience), or due to its relative unpredictability, thereby requiring extra processing resources (expectation-based salience). In their view, the link between prominence perception and attention explains how prominence generation on the speaker's side maps to the perceptual processing at the listener's end. Another approach was proposed by Kakouros and Räsänen (2014a, 2015) who suggested using the low predictability (surprisal) of acoustic prosodic features in speech as a cue for attentional orientation and thereby prominence. They also showed that this correlates well with human perception of prominence in English infant directed speech (Kakouros & Räsänen, in press) and that listeners' cues for prominent words can be altered simply by manipulating the statistical properties of F0 trajectories during a brief pre-test familiarization session (Kakouros & Räsänen, 2016).

In general, stimulus-driven attention and prominence in speech seem to be connected. The most common approach in modeling perceptual attention is to look for unusual changes that take place in specific spatial or temporal contexts, therefore, covering methods that focus on looking for something rare, surprising, or novel (see, e.g., Itti & Baldi, 2009), looking for contrasts (see, e.g., Kakouros, Räsänen, & Laine, 2013), or maximizing the information gain from the input (see, e.g., Bruce & Tsotsos, 2009). Itti and Baldi (2009) have argued that surprisal exists only in the presence of uncertainty that can be described in a relative, subjective manner, based on the expectations of the observer (Itti & Baldi, 2009). Therefore, surprisal can be generally formalized within a probabilistic framework where attentional orientation can be defined as a process that focuses on low probability events given the perceiver's existing probabilistic model of the data. In the context of prosodic prominence, an equivalent cue for attention would be a low probability feature value or feature trajectory in an otherwise predictable context (see also Kakouros & Räsänen, in press).

The probabilistic formulation for prominence also connects to a wider range of phenomena, as frequency and predictability effects have been known to play a fundamental role in models of language production and perception (Jurafsky, 1996; Jurafsky, Bell, Gregory, & Raymond, 2001; Baker & Bradlow, 2009; Bell, Brenier, Gregory, Girand, & Jurafsky, 2009; Watson, Arnold, &

Tanenhaus, 2008). Watson, Arnold, and Tanenhaus (2008), for instance, investigated the effects of predictability and importance on acoustic prominence in language production. Their results showed that both importance and predictability affect the acoustic realization of a word where duration is longer and pitch movement is greater for unpredictable words whereas intensity is greater for important words (Watson, Arnold, & Tanenhaus, 2008). In general, predictable, repeated words seem to be less acoustically prominent than unpredictable or new words (see, e.g., Lam & Watson, 2010). Similar phenomena are also observed in the segmental content of the utterances where, according to van Son and Pols (2003a, 2003b), there is a strong correlation between the redundancy (probability) of a phoneme and the level of acoustic reduction. The probabilistic effects correlating with prominence are not, however, limited only to the frequency of a word's occurrence. Contextual effects from the probabilistic relations between words also affect language production. Therefore, words which are strongly related or predictable from their neighboring words are more likely to be phonologically reduced (Jurafsky, Bell, Gregory, & Raymond, 2001). On the other end, words that are least predictable based on their local context seem to be more likely to carry an accent, thereby being more prominent (Pan & Hirschberg, 2000). Therefore, it seems that frequency, contextual probabilities and, in general, predictability all affect prosodic prominence.

Several theoretical proposals have emerged from the apparent relationship between the predictability of the linguistic elements and their acoustic realization. The *Probabilistic Reduction Hypothesis* (*PRH*) states that word forms are reduced when they have a higher probability (Jurafsky, Bell, Gregory, & Raymond, 2001). In PRH, the probability of a word is conditioned on several aspects such as its context and its syntactic and lexical role. The *Smooth Signal Redundancy Hypothesis* (SSRH) proposed by Aylett and Turk (2004) is based on the relationship between syllable reduction (through durational shortening) and linguistic predictability (see also Turk, 2010, for SSRH over words). SSRH proposes that prosodic prominence is employed in order to manage unpredictable elements in speech, thereby smoothing the information profile of a word. In this framework, language redundancy can be seen as the predictability of a syllable, word, or other linguistic units. Overall, the main claim in SSRH is that the information conveyed by speech should be evenly distributed over time (smooth signal), thereby making speech communication an efficient communication channel. A more recent information-theoretic account is the *Uniform Information Density* (UID) proposed by Jaeger (2006) and Jaeger and Levy (2007). In UID, the central argument is that speakers will plan how to structure a message so that elements with high information value are lengthened and elements with low information are shortened (Jaeger & Levy, 2007). The means to achieve this is through *syntactic reduction* where a speaker may choose to reduce less information-dense sentences by reducing the number of words (manipulating the lexical and syntactic options in the language).

Overall, the link between predictability of the linguistic units in speech has been well established and formalized into several proposals, with a few accounts also attempting to explain the underlying cognitive processes. The existing studies (Kakouros & Räsänen, 2014a; in press) also suggest that predictability is reflected at the level of acoustic prosodic features with low-probability prosodic events correlating with human perception of prominence. Given this background, the current work extends the earlier behavioral findings into an unsupervised algorithm for the automatic detection of sentence prominence in speech that makes use of the (un)predictability of prosodic features, and thereby provides a simple, potentially language-independent approach for automatic prominence labeling. Our aim is not to argue that low-level feature predictability is the sole cue to prominence (see, e.g., Arvaniti, 2009, for discussion on rhythmic constraints on prominence), but to show that it can be used for efficient automatic prominence detection in speech, independently of any labeled training data.

1.2. Earlier work on automatic prominence detection

Several previous efforts have focused on the development of systems for the automatic detection of prominence in speech. Prosodic phenomena, such as sentence prominence, encode higher-level information that is not available in segmental acoustics, thereby making their automatic detection particularly important. Application areas are diverse and include, but are not limited to, natural speech synthesis (see, e.g., Mehrabani, Mishra, & Conkie, 2013; Szaszák, Beke, Olaszy, and Tóth, 2015), automatic speech recognition (ASR) (see, e.g., Chen, Hasegawa-Johnson, Cohen, Borys, Kim, Cole, & Choi, 2006; Ananthakrishnan & Narayanan, 2007), and topics in ASR such as spoken content retrieval (see, e.g., Racca & Jones, 2015; Larson & Jones, 2011), video navigation (see, e.g., Patil, Arsikere, & Deshmukh, 2015), and topic tracking (Guinaudeau & Hirschberg, 2011).

The development of algorithms for the automatic detection of prominence is described by a diverse set of approaches with one typological categorization being that between supervised (see, e.g., Kalinli & Narayanan, 2007), semi-supervised (see, e.g., Jeon & Liu, 2012), and unsupervised methods (see, e.g., Tamburini & Caini, 2005). Majority of the earlier work has focused on supervised methods for learning the statistics connecting typical prosodic features to manual markings of prominence in the data (see, e.g., Rosenberg, Fernandez, & Ramabhadran, 2015; Tamburini, Bertini, & Bertinetto, 2014; Sridhar, Bangalore, & Narayanan, 2008; see also Rosenberg & Hirschberg, 2009). As this requires manually annotated prominence labels, applicability, or at least initial training of such approaches is typically limited to highly resourced languages (see, e.g., Moniz et al., 2014; Rosenberg et al. 2012; Maier et al., 2009). Another limitation that has not been widely addressed is the reliability of the prominence markings. Specifically, it seems that inter-transcriber agreement rates on prominence annotations may vary greatly. Naïve subjects have been shown to have an average agreement rate of kappa of approximately 0.46 (±0.16) (see, Mo, Cole, & Lee, 2008, for American English; Kakouros & Räsänen, 2014b, for British English; You, 2012, for Korean; Smith, 2011, for French; Baumann, 2014, for German), which translates to fair to moderate agreement on the Landis and Koch scale (1977). Trained or expert transcribers following a formal annotation procedure such as a prosodic annotation system like ToBI (Tones and Break Indices) (Silverman et al., 1992) seem to have higher agreement rates with kappa values averaging at approximately 0.72 (±0.05), translating to moderate to substantial agreement (Landis and Koch, 1977) (see, Buhmann, Caspers, van Heuven, Hoekstra, Martens, & Swerts, 2002, for Dutch; Yoon, Chavarria, Cole, & Hasegawa-Johnson, 2004; Breen, Dilley, Kraemer, & Gibson, 2012, for American English; Avanzi, Simon, Goldman, & Auchlin**,** 2010, for French), but still leaving room for ambiguity that can be problematic to many supervised machine learning algorithms. Therefore, one important concern and limitation in the development of supervised systems is the extent to which coders agree on the annotations they apply on speech since this will inevitably affect the ultimate performance the system can reach.

Beyond their limitations, supervised learning approaches provide many options in modeling prominence. As prominence prediction can be posed as a standard machine learning problem involving feature extraction, labeling, training, and classification, various studies are available in the literature addressing the problem at different levels of analysis. Specifically, some approaches suggest the adoption of novel acoustic prosodic features that may be better able to represent and capture relative changes in the signal (see, e.g., Mishra, Sridhar, & Conkie, 2012) while others make use of different combinations of acoustic and lexical/syntactic information (see, e.g., Christodoulides & Avanzi, 2014; Obin, Lacheret-Dujour, & Rodet, 2008, for French; Sridhar, Bangalore, & Narayanan, 2008; Ananthakrishnan & Narayanan, 2008; Imoto, Tsubota, Raux, Kawahara, & Dantsuji, 2002; Minematsu, Kobashikawa, Hirose, & Erickson, 2002; Aylett & Bull, 1998, for English; Arnold, Wagner, & Baayen, 2013, for German; Cutugno, Leone, Ludusan, & Origlia, 2012, for Italian and American English). Large part of the literature has focused on examining the performance of different machine learning algorithms on combinations of acoustic and linguistic feature sets. For instance, Wightman and Ostendorf (1994) used a combination of decision trees to map acoustic prosodic observations to probability distributions and a Markov sequence model of the prosodic labels. Their method achieved prominence detection accuracy at the syllable level of 83%. Wang and Narayanan (2007) used support vector machines also on combinations of acoustic features, reporting precision of 82.1% at the word level whereas Sridhar, Bangalore and Narayanan (2008) used combinations of acoustic and syntactic features in a maximum entropy framework with 86% accuracy at the word level (see also Ananthakrishnan & Narayanan, 2008, for an approach using Gaussian Mixture Models and Neural Networks). Recent approaches place more focus on machine learning methods that can provide better sequence modeling and context representation capabilities (modeling short- and long-term contextual factors). For instance, Rosenberg, Fernandez, and Ramabhadran (2015) used Bidirectional Recurrent Neural Networks (BiRNNs) based on acoustic and lexical features in order to model forward and backward dependencies in the data. They reported accuracy at the word level reaching 89.03%. Based on the same principle, Tamburini, Bertini, and Bertinetto (2014) used various types of probabilistic graphical models (PGMs) such as conditional random fields (CRFs) and conditional neural fields (CNFs) on acoustic features, reporting accuracy of 87.5% at the syllable level (see also, Christodoulides & Avanzi, 2014; Chen, Liu, Yang, & Hu, 2012; Cutugno, Leone, Ludusan, & Origlia, 2012; Obin, Rodet, Lacheret-Dujour, 2009). Although the discussed approaches report accuracies in the range between 80% and 90% at word- or syllable-level, it is not possible to directly compare them as they make use of different speech corpora and potentially slightly different evaluation criteria. However, in general, it seems that the overall binary prominence classification accuracy of the supervised systems lies between 80% to 90%.

Instead of using a priori linguistic information (prominence labels), unsupervised methods typically extract acoustic features directly from the speech signal and compute, for instance, prominence scores using different feature combinations (see, e.g., Tamburini & Caini, 2005; Wang & Narayanan, 2007). For example, Tamburini and Caini (2005) proposed a function where individual acoustic feature values are summed, producing a continuous value of prominence for each syllable. Each syllable score is then compared to its two neighboring syllables and if the center-syllable prominence score is higher, it is marked as prominent. In their experiments, an accuracy of 80.61% was observed at the syllable level. Wang and Narayanan (2007) also proposed a method based on prominence scores using a fusion of acoustic features and reported a precision of 80.0% at the word level. Another unsupervised approach using clustering techniques was proposed by Ananthakrishnan and Narayanan (2006). Their method makes use of a clustering algorithm (such as k-means) to partition the acoustic feature space into two clusters (prominent and non prominent) where they report 77.8% accuracy at the syllable level. Other unsupervised approaches attempt to use biologically inspired methodologies based on the assumption that prominence and attention are connected. Kalinli and Narayanan (2007), for instance, proposed an auditory attention model that attempts to mimic the different processing stages in the central auditory system. Their method extracts a set of multiscale features (features analyzed at different scales, such as intensity and frequency contrast) that are combined into a master saliency map. Local maxima that are above a predefined threshold are then marked as prominent. Their model achieves 75.6% accuracy at the syllable and 78.1% at the word level. Finally, there is also limited research in semi-supervised methods where the overall aim is to use a small amount of labeled prominence data for bootstrapping and then use larger amount of unlabeled data to improve the prominence detection performance (see, e.g., Jeon & Liu, 2012). In all, it seems that unsupervised methods typically reach lower performance levels (75%–80% accuracy) when compared to supervised systems (80%–90%).

In this work, we extend our earlier behavioral findings in Kakouros and Räsänen (in press) to a fully unsupervised prominence detection system. The system is based on the hypothesis that prominence perception correlates with the unpredictability of the prosodic features in speech, and therefore we build an unsupervised model that is able to measure the predictability of the prosody using a simple *n*-gram based "language model" operating on the prosodic features. Although computationally straightforward, the proposed approach produces high agreement with the annotators' prominence markings and allows flexibility with different types of temporal constraints or linguistic units available during the prominence detection task.

2. 3PRO method for automatic prominence detection

We model the prominence detection problem as a binary classification task. The underlying principle in the proposed computational model is to detect prominence in a manner hypothesized to be analogous to human prominence perception. Specifically, the algorithm models the acoustic prosodic feature values and their temporal combinations using *n*-gram statistics (see Fig. 1). Probabilities of the trajectories can be then evaluated on previously unseen utterances where the probabilities are integrated over a fixed-length sliding window or over externally provided linguistic elements (syllables or words), according to their availability. The core principle in determining prominence is then based on finding the points where the prosodic predictability is lowest. Details of the individual components of the algorithm are given below.

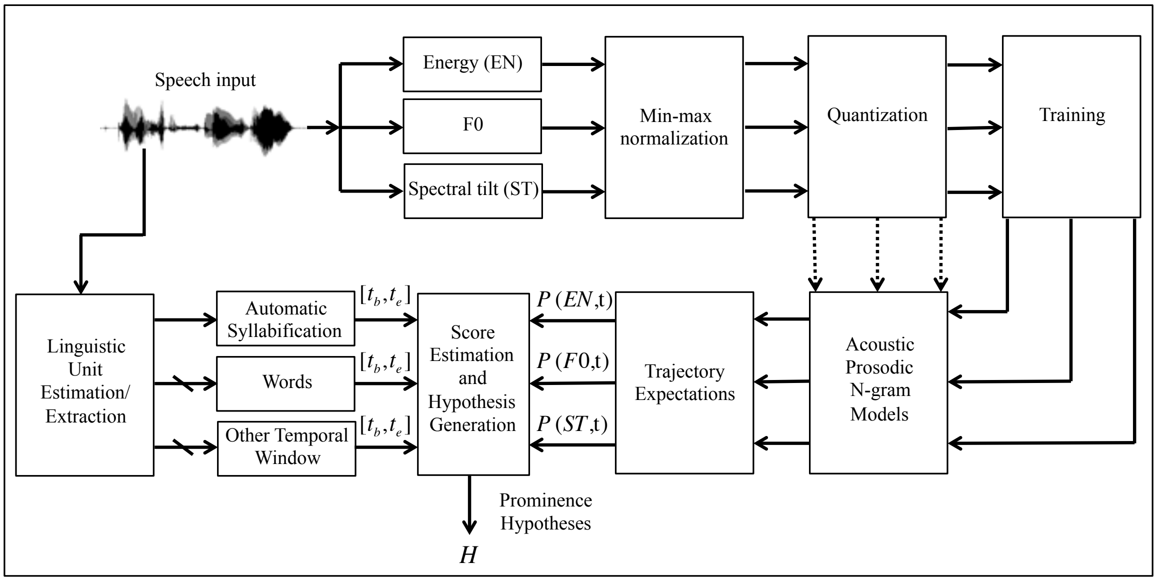

Fig. 1. Overview of the prominence detection algorithm.

2.1. Feature extraction and quantization

Speech data are initially downsampled to 8 kHz. Three acoustic prosodic features, F0, spectral tilt, and energy, are then computed. F0 contours of the voiced segments are extracted from each utterance using the YAAPT algorithm (Zahorian & Hu, 2008) with a 25-ms window and 10-ms step size. The resulting pitch tracks are then linearly interpolated in order to preserve continuity during unvoiced sections. Mel frequency cepstral coefficients (MFCCs) are computed using the same window and step size and the first MFCC is used to represent spectral tilt (see Tsiakoulis, Potamianos, & Dimitriadis, 2010). Finally, signal energy is computed using the same parameters window length as follows:

$$E(t)=\sum_{\tau=-w/2}^{w/2-1}\left|x(t+\tau)\right|^{2}, \qquad (1)$$

where *x* is the speech input and *w* is the length of the analysis window (25 ms).

In order to ensure comparability of the features and reduce the effect of both inter-speaker and intra-speaker variation, energy, F0, and spectral tilt are min-max normalized across each utterance using Eq. (2).

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

In the equation, *f* denotes the feature value at time *t* while max(*f*) and min(*f*) refer to the maximum and minimum values of the feature, respectively, during the given utterance (see, e.g., Imoto et al., 2002). Min-max normalization preserves the relationships between the original feature values but removes information regarding their absolute values, mapping the original feature space to a specific range: $f(t) \in [\min(f)\ \ \max(f)] \rightarrow f'(t) \in [0, 1]$. Effectively, this helps to account for differences in dynamics of the prosody across different talkers and speaking styles, enforcing a prosodic contour for each utterance independently of the absolute magnitude of the features.

Finally, in order to enable discrete probability modeling of the extracted continuous feature values, the prosodic feature values are quantized into *Q* discrete amplitude levels, $f'(t) \rightarrow a_t \in [1, 2, \ldots, Q]$. In the present study, $Q \in [2, 4, 8, 16, 32]$ quantization levels were evaluated using the k-means Linde-Buzo-Gray (LBG) algorithm (Linde, Buzo, & Gray, 1980; see the experiments in section 4).

2.2. Syllable segmentation

The main durational feature in the method is based on syllables. In order to estimate the locations and durations of syllables in an utterance, a signal envelope-based segmentation algorithm proposed in Räsänen, Doyle and Frank (2015) is used as it was shown to compare favorably against other compared methods in a zero-resource speech processing task. The method is based on a harmonic oscillator that is driven by the amplitude envelope of speech. The input envelope is first obtained by filtering the full-wave rectified signal waveform with a low-pass filter that approximates the temporal window of human hearing. The oscillator is then used to smooth this envelope further with time-constraints typical to syllable-rate amplitude modulations in speech. As a result, each sufficiently deep minimum in the oscillator amplitude is interpreted as a syllable boundary while oscillator maxima correspond to nuclei. In the present method, the oscillator was tuned to a center frequency of 4 Hz with a critical damping (Q-factor of 0.5) similarly to the original study (see Räsänen et al., 2015, for details).

2.3. Modeling prosodic trajectories in terms of statistical unpredictability of the prosodic features

The central principle of the 3PRO-algorithm is modeling of acoustic prosodic trajectories in order to capture unpredictable points within a given context. For this purpose, *n*-grams over quantized features are utilized. The probabilities for the *n*-grams are computed from the relative frequencies of different *n*-tuples in the training data according to Eq. (3) where *C* denotes the frequency counts of the discrete n-tuples and ψ the feature in question (e.g., ψ=F0).

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

During prominence detection, features are extracted similarly to training and quantized to *Q* levels. The overall probabilities *P'*(*t*) for *n*-grams across all features are then computed according to Eq. (4) as a sum of feature-specific log-probabilities. The assumption of conditional independence of the studied prosodic features may not hold for all languages and feature combinations, but is the most general way of combining the information without making the approach specific to certain languages.

$$P'(t) = \sum\nolimits_{\psi} \log_{10} \left(P_{\psi}(a_t \mid a_{t-1},\ ...,a_{t-n+1}) \right) \tag{4}$$

The overall surprisal of the features within a time-window is obtained by integrating them with a simple moving average filter of length *L* and changing the sign of the result, leading to the so-called moving average contour (MAC) (Eq. (5)). Thus, each point in MAC reflects a representation of the overall predictability in the immediate surrounding context.

$$MAC(t) = -\sum_{\tau=-L/2}^{L/2-1} P'(t+\tau) \tag{5}$$

As duration is another important correlate of the perceived prominence in speech, with longer durations typically reflecting increased prominence, durational information from the automatic syllabification (section 2.2) is used to modulate the probability contours. More specifically, each time point in MAC is assigned to the syllable enclosing it and the value is modulated according to

$$Y(t) = MAC(t) \times e^{d(t)}\ , \tag{6}$$

where *d*(*t*) is the duration of the syllable enclosing the time-frame *t* (see Figs. 2 and 3).

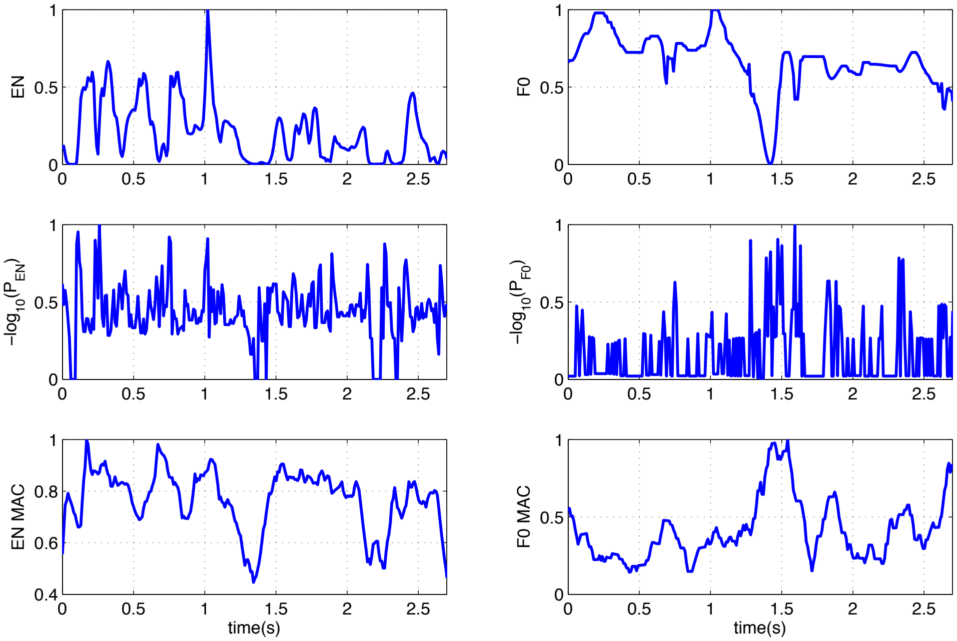

Fig. 2. Example output of the algorithm for a Dutch broadcast excerpt. Top panel: original feature signals. Middle panel: log probabilities for $Q = 16$, $n = 2$ (n-grams). Bottom panel: MAC signals for $L = 200$ ms. All signals are min-max normalized for consistency of presentation.

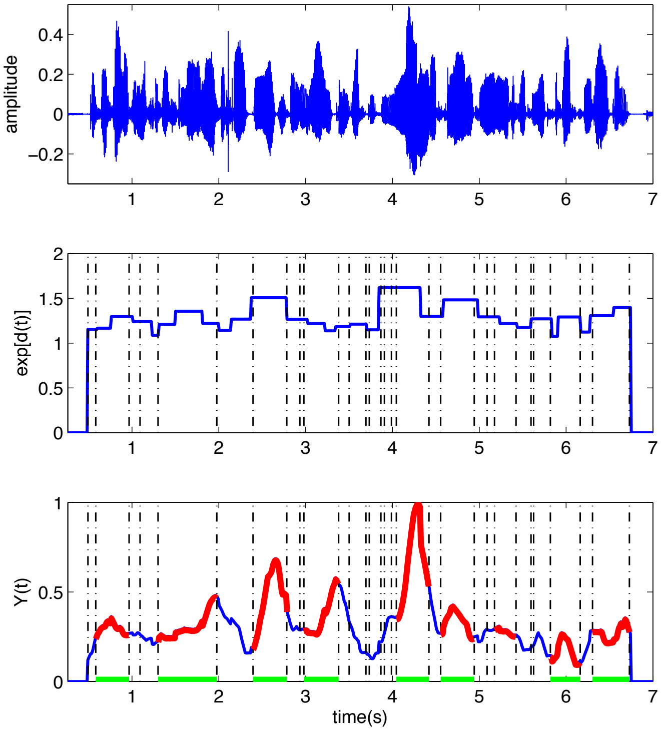

Fig. 3. Example output of the algorithm for a Dutch broadcast excerpt. Top panel: original signal waveform. Middle panel: the syllable signal. Vertical dashed lines denote the word boundaries. Bottom panel: $Y(t)$ for energy and F0, Q = 16, n = 1, L = 200 ms, where the thick red lines mark the word-level human annotated sentence prominence and green the algorithm hypotheses.

2.4. Hypothesis generation without externally provided time-frames

In order to generate prominence hypotheses from $Y(t)$, two different approaches can be used: 1) fully unsupervised estimation of prominent signal regions in continuous time, or 2) determination of prominence within externally provided time windows such as word boundaries.

In the case of the purely unsupervised approach, a simple maxima detection method is employed. First, the $Y(t)$ signals are min-max normalized between [0, 1] (see Eq. (2)). Local maxima are then defined as peaks of $Y(t)$ preceded by a valley with an amplitude smaller than δ of the peak amplitude. In the present study, the value of δ = 0.1 was selected after heuristically searching for the optimal value in the data. All the temporal locations where local maxima occur are set as prominence hypotheses points, $S(t) = 1$, while all other are set to $S(t) = 0$.

In order to generate word-level prominence hypotheses for the experimental evaluation of the system (section 4), each word occurring in the test set was checked for hypothesis points and marked as prominent if one or more prominence markings occur during the word, i.e.,

$$H(w_{ij}) = \begin{cases} 1, & if \sum_t S(t) \geq 1, \quad t_{\text{start}}(i,j) \leq t < t_{\text{end}}(i,j) \\ 0, & otherwise \end{cases} \tag{7}$$

where t_{start} and t_{end} denote the onset and offset times of word i, respectively.

2.5. Hypothesis generation with external time-frames

In scenarios where temporal frames, such as word boundaries, are available, the sliding window integration is replaced by integration of syllable-duration modulated probabilities over the units of interest (Eq. 8). Specifically, in the case of words, the prominence classification $H(w_{i,j})$ for each word i in utterance j is determined based on whether the word-level score $S(w_{i,j})$ falls below a threshold r_i (Eq. (9)) that is defined at the utterance level according to the mean μ and standard deviation σ of the scores across the entire utterance (Eq. (10)), and where hyperparameter λ controls the sensitivity of the prominence detector. Thus, for each word that falls below threshold r_i the corresponding prominence hypothesis for that word $H(w_{i,j})$ is set to 1. In the experiments, hyperparameter λ is varied between [-2,2] in order to evaluate algorithm performance for difference threshold levels.

$$S(w_{ij}) = \sum_{t=t_{\text{start}}}^{t_{\text{end}}} P'(t) \times e^{d(t)} \tag{8}$$

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

$$r_i = \mu_i - \sigma_i \lambda \tag{10}$$

2.6. Baseline systems

In order to evaluate the performance of the proposed prominence detection method, the 3PRO output was compared to a number of baseline setups. Specifically, the first is a purely random baseline approach (RBA) that randomly selects *N* words as prominent in each utterance where *N* equals to the number of word hypotheses generated by the 3PRO-algorithm for the same signal. The second method uses raw feature maxima (RFM) in order to mark words as prominent. In particular, for each utterance, the *N* words with the highest absolute feature values are selected (e.g., the *N* words with the highest pitch), *N* again equaling to the number of word hypotheses generated by the 3PRO.

In addition, in order to facilitate comparison to the ideal supervised case, standard supervised classification baselines were computed using the k-nearest-neighbor (kNN) classifier, support vector machines (SVMs), and conditional random fields (CRFs), using the manually labeled section of the data used in the experiments. The first two represent context-independent classifiers where prominence classification for a word is independent of the neighboring classification decisions (but see the features below), whereas CRFs take the entire chain of classifications into account in hypothesis generation.

All hyperparameters of kNN, SVMs, and CRFs were optimized for maximal performance on the test data of the CGN corpus. Classification performance for kNN was computed for all values of k and the best result was chosen as the reference. For SVMs, radial basis function kernel was used. The scaling factor σ and box-constraint C of the SVMs were optimized by first using a subsampling scheme to find an initial estimate σ_{init} and then using an exhaustive grid-search across $\sigma = [0.001, 0.01, \ldots, 1000]\times\sigma_{\text{init}}$ and $C = [0.001, 0.01, \ldots, 1000]$. Linear chain CRFs were trained using belief propagation and L2 regularization with maximum number of iterations set to 100, as this was manually verified to lead to convergence of the training error. CRF regularization parameter of the regularization term $-\theta^2/2\sigma_{\text{reg}}^2$ was set to $\sigma_{\text{reg}} = 1$ after manually experimenting over a range of values.

The basic features used in the supervised baselines were calculated separately for each word, including max, min, mean, variance, and difference (x_{onset}-x_{offset}) of energy, tilt and pitch during the words, word log-duration, and the number of syllabic nuclei in the words extracted as envelope maxima from the method described in Räsänen et al. (2015). In addition, in order to account for the word context also in kNNs and SVMs, we followed Rosenberg et al. (2012) by including the number of syllabic nuclei, mean and max pitch, and mean and max energy of the two preceding and two following words in the set of features, leading to a total set of 37 unique features per word token.

We also studied the inclusion of the features proposed for prominence classification by Mishra et al. (2012), including area under F0 curve (AFC), energy-F0-integral (EFI), voiced-to-unvoiced ratio (VUR), average difference between low and high frequency components (DLH), and F0 peak/valley amplitude and location (FAMP & FLOC) (see Mishra et al., 2012, for details) to our feature set. However, these did not lead to further improvements with respect to the above reported set of features and were therefore excluded from the final experiments. All features were mean and variance normalized across each recording before inclusion in the classification, as this was found to improve generalization across recording types especially on C-PROM where many different types of recordings are included.

3. Data and evaluation

3.1. Data

3.1.1. Spoken Dutch Corpus

The Spoken Dutch Corpus (Corpus Gesproken Nederlands; CGN) was used in order to evaluate the algorithm's performance for Dutch continuous speech (Oostdijk et al., 2002). CGN is a database of contemporary standard Dutch as spoken by adults in The Netherlands and Flanders. It contains nearly 9 millions words (800 hours of speech), of which approximately two thirds originate from The Netherlands and one third from Flanders. The database contains several manually generated or verified annotations such as phonetic transcriptions, word level alignment, and prosodic annotations (see Duchateau, Ceyssens, & Van Hamme, 2004, for a more detailed description). In the present experiments, the Dutch news broadcast ("*component k*") section of the corpus was used, consisting of 5088 news broadcasts (≈27.4 hours of speech data) spoken by 29 speakers (22 male and 7 female) and containing a total of 285298 words. The prosodically annotated subset of the section consists of 134 news broadcasts spoken by 10 different speakers (9 male and 1 female) and contains a total of 7438 words (≈44.3 minutes of speech data). Each sentence in this subset was hand labeled by two trained annotators (see Buhmann, Caspers, van Heuven, Hoekstra, Martens, & Swerts, 2002, for a description of the annotation process). In the experiments, the full broadcast section plus nine talkers from the annotated section were always used for training of the system while the remaining talker from the annotated subset was used for evaluation, leading to a 10-fold evaluation procedure. As for the supervised baselines, the same 10-fold evaluation was used, but training only with the labeled data from nine speakers available for each fold.

3.1.2. C-PROM

We also evaluated our system on French speech by using the C-PROM corpus (Avanzi, Simon, Goldman, & Auchlin, 2010), a corpus specifically annotated for prominence studies. C-PROM contains different regional varieties of spoken French (Belgian, Swiss, and metropolitan French) as well as various discourse genres with multiple levels of annotations. The corpus includes 24 recordings with 70 minutes of speech produced by 28 speakers (12 female and 16 male) and with 7 different speaking styles (ranging from high to low degrees of formality), totaling to 13184 words. The corpus contains phone, syllable, and word level transcriptions along with syllable-level prominence labels annotated by two expert phoneticians. The prominence labeling is based on a consensual annotation where the two annotators discussed and resolved potential differences in the coding, resulting in a single set of prominence labels for the data (see Avanzi, Goldman, Lacheret-Dujour, Simon, & Auchlin, 2007, for more details). In the experiments, one recording is always used for testing while the remaining 23 recordings are used for training, resulting in a 24-fold evaluation procedure.

3.2. Evaluation

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

$$\mathrm{RCL} = tp\,/\,(tp + fn) \tag{11}$$

$$\mathrm{PRC} = tp\,/\,(tp + fp) \tag{12}$$

$$\mathrm{F} = (2 \times \mathrm{PRC} \times \mathrm{RCL})/(\mathrm{PRC} + \mathrm{RCL}) \tag{13}$$

$$\mathrm{ACC} = (tp + tn)/(tp + fp + fn + tn) \tag{14}$$

where *tp* denotes the true positives, *tn* the true negatives, *fp* the false positives, and *fn* the false negatives. Fleiss kappa (Fleiss, 1971) was also used as a measure of the reliability of agreement between the algorithm and the annotators' judgments as it allows direct comparison to the typical agreement between human annotators (see, e.g., Mo, Cole, & Lee, 2008; You, 2012). Overall, Fleiss kappa measures the degree of agreement between two or more annotators on a nominal scale of $\kappa \in [-1,1]$ and yields $\kappa = 0$ if the number of agreements is equal to what is expected based on chance-level co-occurrences in the data and $\kappa = 1$ if all annotators fully agree. In this work, Fleiss kappa was measured at the word-level as also the data in CGN provided word-level annotations. Therefore, for each word occurring in the test set, a binary decision between non-prominent and prominent was considered. For C-PROM, the syllable-level prominence labels were aligned with the word-level transcriptions in order to provide word-level binary prominence annotations.

For CGN, results were computed by taking the mean pairwise agreement between the algorithm and the annotators (kappa) or the mean accuracy and F-score with respect to the two different annotations. This scenario is referred to as *true human* (TH) reference. In addition, results were also computed using a reference where all words that either or both of the annotators had labeled as prominent were marked as prominent. This scenario is referred to as *broad annotation* (BA) reference. For C-PROM corpus, since there was only one set of reference labels for the data (see section 3.1.2.), the agreement between the algorithm and the word-level labels was measured directly.

3.3. Annotation data analysis

In order to understand the characteristics of the reference annotation, we computed basic summary statistics from the data. First, for the CGN corpus, the overall word-level agreement between the two annotators on the presence or absence of prominence was found to be κ = 0.68. Both of the annotators marked, on average, 35.4% of the words as prominent (from the total set of 7438 words) while they both agree on the presence or absence of prominence on 85.5% of the words. In the case of the BA reference, a total of 42.6% of the words are marked as prominent where the mean agreement between the annotators and combined BA reference is 92.7% (κ = 0.85), providing an approximate value for the performance ceiling achievable by any automatic system. Overall, the inter-annotator agreement rates are typical for prominence annotation. For instance, Bushmann et al. (2002) report inter-annotator agreements ranging between 0.58 and 0.72 for a similar prominence labeling task in a subset of CGN. Similarly, in another experiment carried out by Streefkerk et al. (1997) on the Dutch polyphone corpus, kappa values across naïve annotators were found to range between 0.45 and 0.6. As for the C-PROM corpus, the authors of the corpus report inter-annotator agreement between the two transcribers to be κ = 0.77 (Avanzi et al., 2010). In C-PROM, 29% of the words are marked as prominent (3825 words) – see also Rosenberg et al. (2012) for a similar study on C-PROM on word-level prominence.

4. Experiments

The prominence detection algorithm was tested in two basic experiments on Dutch and French: The first experiment investigates the performance and optimal parameters for the purely unsupervised approach when the underlying linguistic units are not known. The second experiment tests the performance of the system when word boundaries are available during prominence detection (e.g., originating from a parallel ASR system or manual word-level transcription).

4.1. 3PRO without externally provided time-frames

4.1.1. Spoken Dutch Corpus

The first experiment for CGN corpus was run in a cross-validation setup where data from 28 speakers were used for training and 1 for testing. The analysis was limited to *n*-gram orders of *n* = 1, 2, 3, 4, and 5 as modeling with higher orders would result in very sparse statistics. The experiment was repeated for $Q \in \{2, 4, 8, 16, 32\}$ and separately for the three features (energy, F0, spectral tilt) and their combinations. After feature extraction and quantization of the test data, the probabilities of all possible feature combinations were evaluated over temporally varying windows of $L \in [10, 1000]$ ms with steps of 10 ms in order to investigate the optimal value of *L* in the task. Table 1 summarizes all the features and their combinations used in the experiments.

Table 1. Features and feature combinations used in the experiments.

Feature	Description
EN	Energy
F0	Fundamental frequency
ST	Spectral tilt
EN+F0	Energy and fundamental frequency
EN+ST	Energy and spectral tilt
F0+ST	Fundamental frequency and spectral tilt
EN+F0+ST	Energy, fundamental frequency and spectral tilt

Fig. 4 presents the results for Q = 16 and n = 1 for all individual features and the best feature combination of energy with F0. The quantization level was selected as a compromise between too coarse and too fine clustering for an initial analysis of the effect of the temporal integration window size. Similarly, $n = 1$ is the simplest expectation estimation that reflects the marginal probabilities of the discrete amplitude levels. Based on the results, it seems that temporal integration windows at or around 200-ms give the best overall performance for all individual features and their combinations. All other combinations were also tested with performance inferior to that of energy with F0. Specifically, for $L = 200$ ms, the mean pairwise agreement for energy, F0, ST, and EN+F0, was 0.47, 0.53, 0.3, and 0.56, respectively. In the case of the BA reference, the results for the same features were kappa of 0.56, 0.50, 0.34, and 0.62 respectively. We also found that the maximum peak for energy took place at 260 ms ($\kappa_{TH} = 0.48$, $\kappa_{BA} = 0.56$) while the peak for F0 was located at 130 ms ($\kappa_{TH} = 0.56$, $\kappa_{BA} = 0.56$). For ST and EN+F0 the maximum peaks took place exactly at 200 ms.

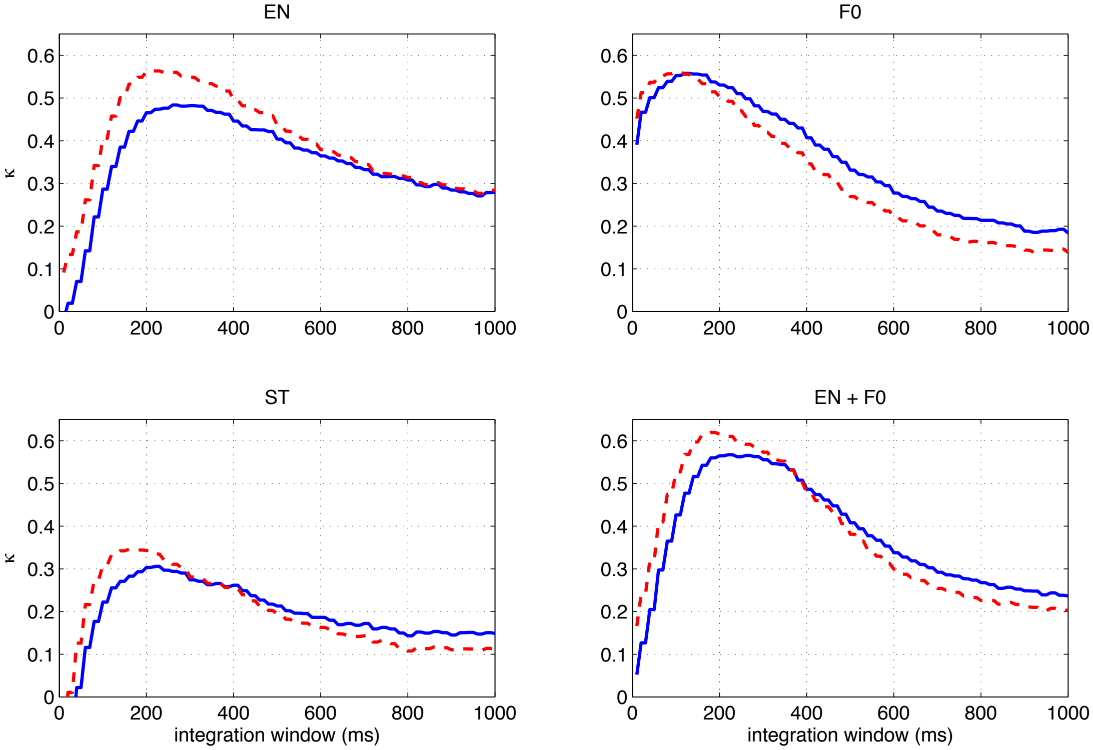

Fig. 4. Effect of temporal integration window size (ms) on algorithm performance for Q = 16 and n = 1. Blue solid line represents the mean pairwise agreement between the algorithm and the annotators while the red dashed line represents the agreement between the BA reference and the algorithm.

Next, the impact of the number of quantization levels on algorithm performance was evaluated on CGN. In this case, Q = 8 and 16 were found to be the optimal with performance being almost equal between the two and deteriorating for codebooks smaller than 8 and larger than 16 levels (see also Table 2). Specifically, for n = 1 and Q = 8, $\kappa_{\mathrm{BA,EN+F0}} = 0.64$ ($\mathrm{ACC}_{\mathrm{BA,EN+F0}} = 82.3\%$ and $\mathrm{F}_{\mathrm{BA,EN+F0}} = 79.9\%$) and $\kappa_{\mathrm{TH,EN+F0}} = 0.58$ ($\mathrm{ACC}_{\mathrm{TH,EN+F0}} = 80.2\%$ and $\mathrm{F}_{\mathrm{TH,EN+F0}} = 74.9\%$), whereas for n = 1 and Q = 16, $\kappa_{\mathrm{BA,EN+F0}} = 0.62$ ($\mathrm{ACC}_{\mathrm{BA,EN+F0}} = 81.2\%$ and $\mathrm{F}_{\mathrm{BA,EN+F0}} = 79.1\%$) and $\kappa_{\mathrm{TH,EN+F0}} = 0.57$ ($\mathrm{ACC}_{\mathrm{TH,EN+F0}} = 79.7\%$ and $\mathrm{F}_{\mathrm{TH,EN+F0}} = 73.9\%$). Therefore, Q = 8 was used in the remaining experiments. Finally, in terms of the n-gram order, n = 1 was found to be the optimal with increasing orders slightly deteriorating overall performance (Table 2). It was also observed that with orders of $n > 1$, the optimal integration window size L was approximately at 400 ms for most features and feature combinations.

Fig. 5 presents the results for Q = 8, averaged across n-gram orders from 1 to 5 for the best performing features and their combination using the BA reference and plotted together with the RFM and RBA baselines. In particular, raw feature maxima values at the word level for energy seem to provide a good cue for prominence reaching agreement levels up to $\kappa_{\mathrm{BA,RFM,EN}} = 0.47$. However, algorithm performance is substantially higher reaching kappa values up to 0.56. Random selection of words in the RBA baseline leads to negative kappa values ($\kappa_{\mathrm{BA,RBA,EN}} < 0$) indicating that there is no agreement with the reference annotation. F0 RFM reaches only $\kappa_{\mathrm{BA,RFM,F0}} = 0.32$ at best, showing that purely picking F0 word-level maxima may not be as indicative of prominence as opposed to $\kappa_{BA,\ F0} = 0.53$ from the algorithm. Finally, the RFM combination of energy together with F0 reaches only $\kappa_{BA,RFM,EN+F0} = 0.33$ indicating that combining the raw feature values for maxima picking is not an effective strategy as it deteriorates the overall agreement when compared to energy alone.

In all, the fully unsupervised system performs the best for $Q = 8$, $n = 1$, for window integration length of 200 ms, and for the feature combination of energy together with F0. With this combination, the system can reach high agreement with $\kappa_{BA} = 0.64$ and accuracy of 82.3% providing good overall performance, adding to that of existing unsupervised systems. For instance, Kalinli and Narayanan (2007) report accuracy of 78.1% at the word level and Ananthakrishnan and Narayanan (2008) 77.8% accuracy at the syllable level. Even though the results are not directly comparable due to the use of different speech corpora, they are indicative of the overall classification performance.

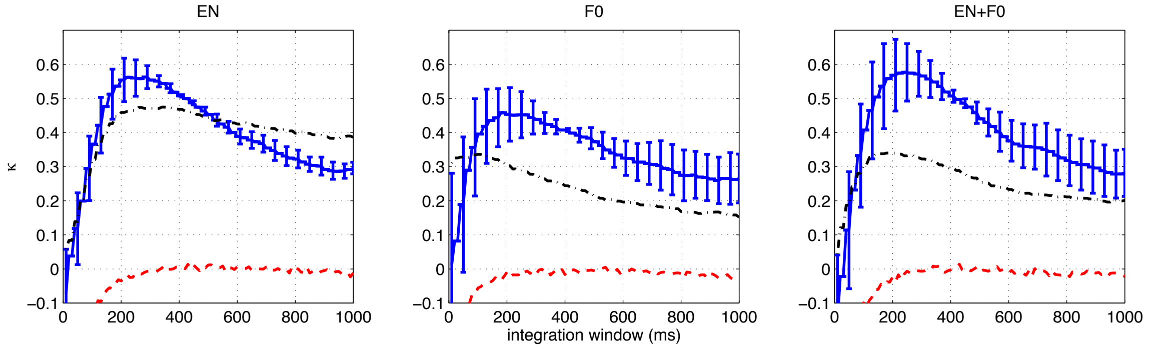

Fig. 5. Fleiss kappa for energy, F0 and their combination pooled over n-gram orders of 1 to 5, for $Q = 8$, and using the BA reference (blue solid line). Black dashed-dotted line represents the feature maxima baseline (RFM) whereas the red dashed line represents the random baseline (RBA). Vertical bars denote one standard deviation across n-gram orders.

Table 2. N-gram performance for $n = 1$, 2, and 3, $Q = 8$, for the best integration window size, for the BA and TH reference. Values in bold indicate the best results for each measure.

Features	κ		ACC		PRC		RCL		F	
n = 1, **L = 200 ms**	*BA*	*TH*	*BA*	*TH*	*BA*	*TH*	*BA*	*TH*	*BA*	*TH*
EN	0.56	0.45	0.78	0.73	0.68	0.57	**0.90**	**0.91**	0.78	0.71
F0	0.53	0.55	0.78	**0.80**	**0.80**	**0.72**	0.64	0.70	0.71	0.71
ST	0.33	0.29	0.67	0.65	0.59	0.51	0.69	0.71	0.64	0.59
EN+F0	**0.64**	**0.58**	**0.82**	**0.80**	0.77	0.67	0.83	0.86	**0.80**	**0.75**
EN+ST	0.48	0.40	0.74	0.71	0.65	0.57	0.82	0.84	0.73	0.67
ST+F0	0.46	0.44	0.74	0.73	0.68	0.60	0.71	0.74	0.70	0.66
EN+F0+ST	0.55	0.51	0.78	0.76	0.72	0.63	0.77	0.81	0.75	0.71
n = 2, **L = 400 ms**										
EN	0.50	0.44	0.75	0.72	0.68	0.58	0.79	0.81	0.73	0.68
F0	0.36	0.32	0.68	0.67	0.61	0.52	0.71	0.73	0.66	0.61
ST	0.35	0.29	0.68	0.65	0.60	0.51	0.73	0.74	0.66	0.60
EN+F0	0.50	0.43	0.75	0.72	0.68	0.58	0.77	0.79	0.72	0.67
EN+ST	0.44	0.39	0.73	0.71	0.66	0.57	0.72	0.74	0.69	0.64
ST+F0	0.37	0.32	0.69	0.67	0.62	0.52	0.70	0.72	0.66	0.60
EN+F0+ST	0.44	0.39	0.73	0.71	0.67	0.57	0.71	0.73	0.69	0.64
n = 3, **L = 400 ms**										
EN	0.49	0.43	0.75	0.72	0.68	0.58	0.77	0.79	0.72	0.67
F0	0.36	0.32	0.68	0.67	0.62	0.53	0.66	0.68	0.64	0.60
ST	0.34	0.30	0.67	0.66	0.60	0.51	0.69	0.70	0.64	0.59
EN+F0	0.46	0.40	0.73	0.71	0.67	0.57	0.74	0.76	0.70	0.65
EN+ST	0.43	0.38	0.72	0.70	0.66	0.56	0.70	0.72	0.68	0.63
ST+F0	0.35	0.31	0.68	0.66	0.61	0.52	0.66	0.68	0.64	0.59
EN+F0+ST	0.42	0.38	0.72	0.70	0.66	0.56	0.69	0.71	0.67	0.63

4.1.2. C-PROM

For C-PROM corpus, the first experiment was run in a cross-validation setup where data from 23 recordings (26–27 speakers) were used for training and 1 for testing (1–2 speakers), leading to a total of 24 folds. The procedure was otherwise the same with CGN (section 4.1.1.) using the features and feature combinations listed on Table 1 for the C-PROM data and where the probabilities of all possible feature combinations were computed over temporally varying windows of $L \in [10, 1000]$ ms with steps of 10 ms. Fig. 6 presents the results for $Q = 16$ and $n = 1$ for all individual features and the best feature combination of energy with F0. The best overall performance for the individual features of energy and ST was attained for temporal integration windows of 150-ms, wheras, for F0, the optimal window size was 50 ms. For the combined features' performance, the best result was achieved with energy and F0 using an integration window of 150 ms. Specifically, for L = 150 ms, the agreements for energy, F0, ST, and EN+F0, were 0.43, 0.38, 0.32, and 0.49, respectively. All other feature combinations were also tested with the performance being inferior to that of energy with F0.

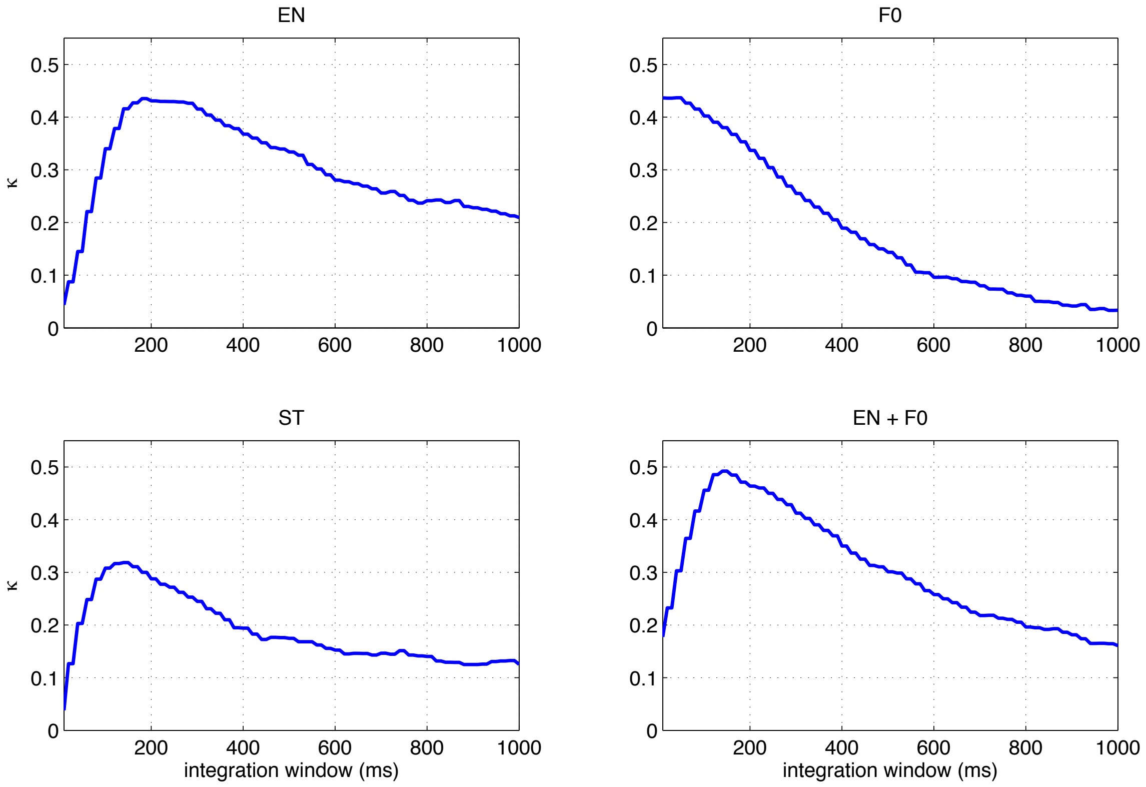

Fig. 6. Effect of temporal integration window size (ms) on fully unsupervised algorithm performance for Q = 16 and n = 1. Blue solid line represents the agreement between the algorithm and the C-PROM annotation reference.

Next, the effect of the number of discrete amplitude levels Q on the algorithm's performance was evaluated. Similar to the results for CGN, Q = 8 and 16 were found to be the optimal partitions, with $Q < 8$ producing poor performance while $Q > 16$ led to slightly deteriorating performance, where, for instance, for $n = 1$ and $Q = 8$, $\kappa_{EN+F0} = 0.49$ ($ACC_{EN+F0} = 77.5\%$ and $F_{EN+F0} = 66.02\%$) and $n = 1$ and $Q = 16$, $\kappa_{EN+F0} = 0.48$ ($ACC_{EN+F0} = 77.5\%$ and $F_{EN+F0} = 63.88\%$). Thus, $Q = 8$ was used in the remaining experiments. Finally, the effect of different n-gram orders was investigated and $n = 1$ was found to produce the best performance with $n > 1$ generating slightly deteriorating performance – see also Table 3.

Fig. 7 presents the results for $Q = 8$, averaged across n-gram orders and for the best performing features and their combination plotted together with the RFM and RBA baselines. Raw feature maxima values produced lower overall performance when compared to the individual and combined features' of the proposed probabilistic approach. In particular, raw feature maxima at the word level for energy, F0, and their combination reached, at best, a performance of $\kappa = 0.39$, 0.26, and 0.37 respectively, while, the corresponding probabilistic frame-based integration reached $\kappa = 0.44$, 0.44, and 0.49. Additionally, a random selection of words in the RBA baseline led to negative kappa values ($\kappa<0$), indicating no agreement with the reference annotation. Therefore, even though the absolute feature values seem to give an indication of the prominent words, the performance is substantially lower when compared to the probabilistic approach. Also, as can be seen in Fig. 7, the best overall integration window size seems to be at approximately 200 ms, as was observed also in Fig. 5 for the CGN corpus.

In all, the results for the fully unsupervised system for the C-PROM corpus are in line with those for CGN. The best performance was achieved for $Q = 8$, $n = 1$, for an integration window of 150 ms, and for the feature combination of energy together with F0. Based on this setup, the system was able to reach good overall performance with $\kappa = 0.49$ and accuracy of 77.5%, a result close to that of other unsupervised approaches (78.1%, Kalinli and Narayanan, 2007).

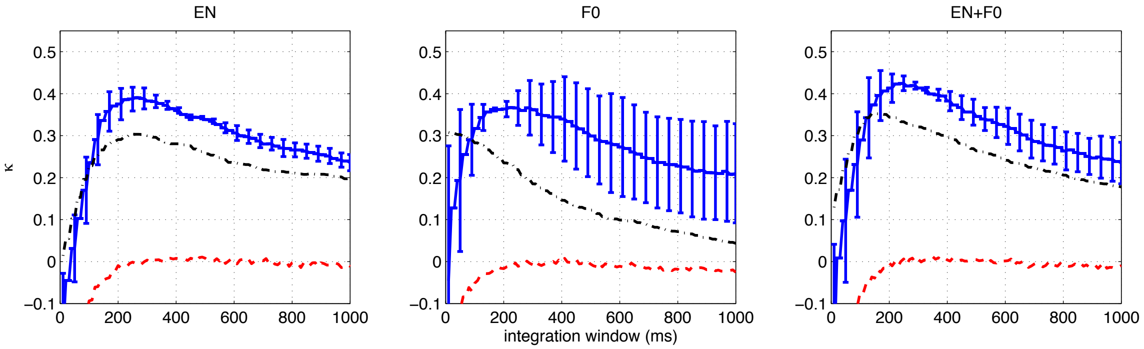

Fig. 7. Fleiss kappa for energy, F0 and their combination pooled over *n*-gram orders of 1 to 4, for *Q* = 8 for C-PROM. Black dashed-dotted line represents the feature maxima baseline (RFM) whereas the red dashed line represents the random baseline (RBA). Vertical bars denote one standard deviation across *n*-gram orders.

Table 3. *N*-gram performance for *n* = 1, 2, and 3, *Q* = 8, for the best integration window size, for C-PROM. Values in bold indicate the best results for each measure.

Features	κ	ACC	PRC	RCL	F
***n* = 1,** ***L* = 150 ms**					
EN	0.42	0.73	0.52	**0.81**	0.63
F0	0.38	0.77	**0.66**	0.44	0.53
ST	0.32	0.70	0.49	0.59	0.54
EN+F0	**0.49**	**0.78**	0.60	0.76	**0.66**
EN+ST	0.38	0.73	0.52	0.68	0.59
ST+F0	0.39	0.74	0.55	0.60	0.57
EN+F0+ST	0.44	0.75	0.56	0.68	0.62
***n* = 2,** ***L* = 400 ms**					
EN	0.36	0.73	0.53	0.60	0.56
F0	0.39	0.75	0.56	0.59	0.57
ST	0.28	0.69	0.47	0.53	0.50
EN+F0	0.41	0.75	0.56	0.61	0.58
EN+ST	0.34	0.72	0.52	0.55	0.53
ST+F0	0.34	0.72	0.52	0.58	0.55
EN+F0+ST	0.37	0.74	0.54	0.57	0.56
***n* = 3,** ***L* = 400 ms**					
EN	0.37	0.73	0.53	0.62	0.57
F0	0.39	0.74	0.54	0.61	0.58

ST	0.28	0.69	0.47	0.54	0.50
EN+F0	0.40	**0.75**	0.56	0.62	0.59
EN+ST	0.33	0.72	0.51	0.56	0.53
ST+F0	0.34	0.72	0.51	0.58	0.54
EN+F0+ST	0.37	0.74	0.54	0.58	0.56

4.2. 3PRO with word-based time frames

4.2.1. Spoken Dutch Corpus

The second experiment for CGN was run similarly to experiment 1 (section 4.1.1.) using a 10-fold evaluation procedure utilizing the same statistical models for *n*-grams for energy, F0, and spectral tilt. During testing, after feature extraction and quantization, the instantaneous probabilities for each utterance and for all feature combinations were computed (see Table 1). Scores were then calculated for each word in each utterance by integrating the prosodic feature probabilities over the word durations (see section 2.5) and prominence hypotheses were generated by thresholding the word-specific score values. In order to evaluate performance for different threshold levels, hyperparameter λ was varied between [-2,2] with steps of 0.05. Fig. 8 presents the results for the three individual features and the best performing feature combination of energy together with F0 for the optimal parameterization. Specifically, for $Q = 16$, $n = 2$, $\lambda = -0.15$ and for the BA reference EN+F0 reached $\kappa_{BA} = 0.72$, ACC = 85.5% and F = 83.6%. The corresponding results for the TH reference were $\kappa_{TH} = 0.63$, ACC = 82.5% and F = 77.4%. Individual features seem to perform also well, with $\kappa_{BA,EN} = 0.68$, $\kappa_{BA,F0} = 0.62$, and $\kappa_{BA,ST} = 0.65$ (see also Table 4). All possible parameterizations for Q and n were also tested with performance being worse for the lowest quantization level ($Q = 2$) while using unigrams. Overall, the performance keeps increasing with an increasing number of amplitude levels up to $Q = 32$ with $Q = 16$ and 32 being the best and approximately equally performing codebooks. As for the *n*-gram order, $n = 2$ and 3 seem to provide the best performance. Higher orders also produce good results with, however, decreasing performance due to the increasing sparsity of the probability space (see also Table 4). Therefore, $Q = 16$ and $n = 2$ can be seen as the best parameters for the unsupervised word duration-based system in the present dataset.

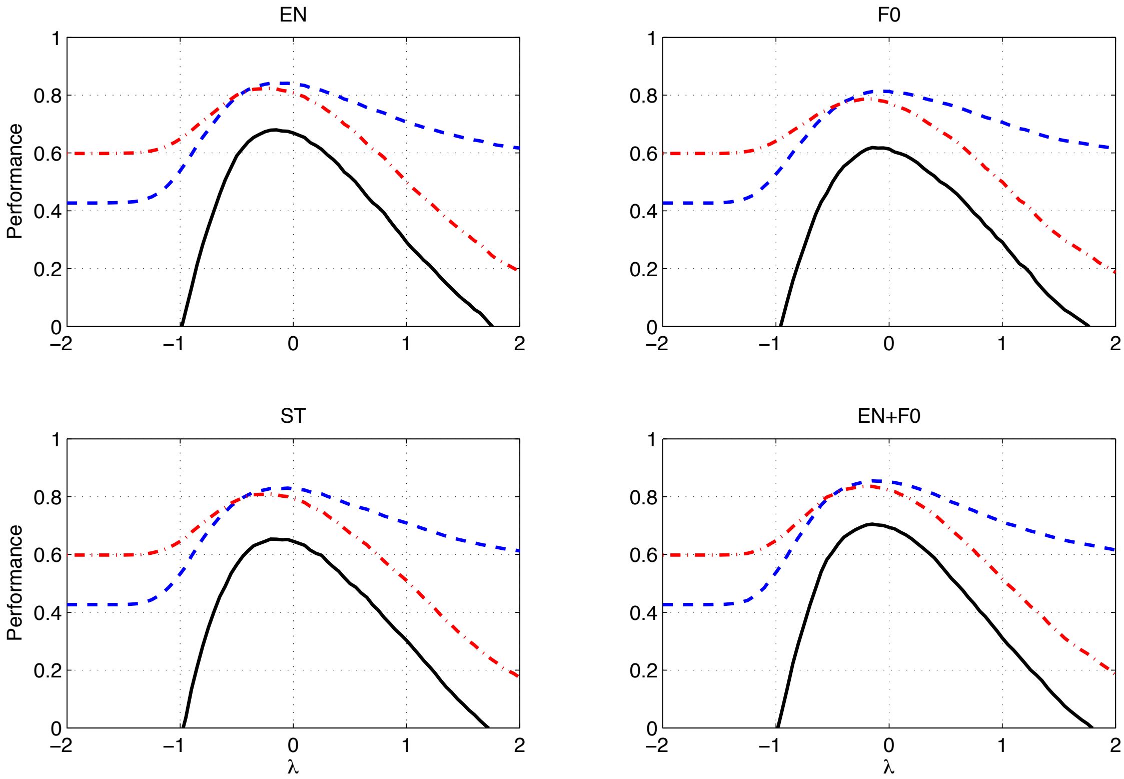

Fig. 8. Performance based on probability integration over word durations for individual features and their best combination for the BA reference, for $Q = 16$ and $n = 2$, and as a function of the detection threshold λ. Blue dashed line represents accuracy, red dash-dotted line represents F-score, and black solid line represents the Fleiss kappa agreement.

Since all features seem to perform relatively well in the word-level decoding, a duration-based baseline model was also evaluated using the same dynamic thresholding as in Eq. (9) but operating on word durations instead of word scores. Word duration is known to be a very important correlate of prominence (see, e.g., Moubayed, Ananthakrishnan, & Enflo, 2010) and this is also seen in the present results, where, for $\lambda = -0.15$, $\kappa_{\mathrm{BA}} = 0.64$ and $\kappa_{\mathrm{TH}} = 0.57$ are obtained. Overall, when integrating probabilities at the word level, performance is substantially higher when compared to the integration over fixed-length windows, and this is largely supported by the access to lexically-constrained durational information.

Table 4. Performance of the word-level integration tests for n = 1, 2, and 3, Q = 16, λ = -0.15, for the BA and TH reference. Values in bold indicate the best results for each measure.

Features	κ		ACC		PRC		RCL		F	
***n* = 1**	BA	TH	BA	TH	BA	TH	BA	TH	BA	TH
EN	0.68	0.59	0.84	0.80	0.80	0.67	0.84	0.86	0.82	0.75
F0	0.47	0.48	0.74	0.75	0.73	0.65	0.65	0.69	0.68	0.67
ST	0.55	0.49	0.78	0.75	0.73	0.62	0.77	0.79	0.75	0.69
EN+F0	0.66	0.60	0.83	0.81	0.80	0.68	0.81	0.84	0.80	0.75
EN+ST	0.67	0.58	0.84	0.80	0.80	0.67	0.84	0.86	0.82	0.75
ST+F0	0.62	0.57	0.81	0.79	0.78	0.67	0.78	0.82	0.78	0.74
EN+F0+ST	0.68	0.61	0.84	0.81	0.80	0.69	0.84	0.86	0.82	0.76
***n* = 2**										
EN	0.69	0.60	0.85	0.81	0.80	0.69	0.85	0.84	0.83	0.76
F0	0.67	0.58	0.82	0.80	0.77	0.68	0.82	0.80	0.79	0.73
ST	0.60	0.58	0.84	0.80	0.79	0.68	0.85	0.82	0.82	0.74
EN+F0	**0.72**	**0.63**	**0.86**	**0.83**	0.80	**0.71**	**0.86**	0.85	**0.84**	**0.77**
EN+ST	0.69	0.61	0.85	0.81	0.80	0.70	0.85	0.84	0.83	0.76
ST+F0	0.68	0.60	0.84	0.81	0.80	0.69	0.85	0.83	0.82	0.76
EN+F0+ST	0.70	0.62	0.85	0.82	0.80	0.70	**0.86**	0.85	0.83	**0.77**
***n* = 3**										
EN	0.69	0.60	0.85	0.81	0.80	0.68	0.85	0.87	0.83	0.76
F0	0.59	0.52	0.80	0.77	0.75	0.64	0.79	0.81	0.77	0.72
ST	0.66	0.58	0.83	0.80	0.79	0.67	0.84	0.86	0.81	0.75
EN+F0	0.70	0.61	0.85	0.81	**0.81**	0.69	0.85	**0.88**	0.83	**0.77**
EN+ST	0.69	0.61	0.85	0.81	**0.81**	0.68	0.85	0.87	0.83	0.76
ST+F0	0.67	0.59	0.84	0.80	0.79	0.67	0.84	0.86	0.81	0.75
EN+F0+ST	0.70	0.61	0.85	0.81	**0.81**	0.68	0.85	0.87	0.83	**0.77**

4.2.2. C-PROM

The second experiment for C-PROM corpus was run based on the same procedure as for the CGN corpus (see section 4.2.1.) but using a 24-fold evaluation process utilizing the same statistical *n*-grams models for energy, F0, and spectral tilt as in experiment 1 (section 4.1.2.). Similarly to CGN, the performance for different thresholds λ was evaluated in the range [-2, 2] with steps of

0.05. Fig. 9 presents the results for the individual features as well as for the best feature combination of energy and F0. Specifically, for Q = 16, n = 2, and λ = 0.25, EN+F0 reached κ = 0.58, ACC = 82.3% and F = 70.7%. Individual features performed also well with κ_{EN} = 0.52, κ_{F0} = 0.55, and κ_{ST} = 0.50 (see also Table 5). All possible parameterizations for Q and n were also tested and the best performing partitions were found to be for Q = 16 and 32 while the worst for Q = 2 and 4. As for the n-gram order, n = 1 was the worst performing length for the feature sequences whereas for n > 2 performance was gradually deteriorating with increasing n-gram orders. Thus, as in the case of CGN, the best performing parameterizations for the unsupervised word duration-based system are for Q = 16 and n = 2. Finally, we evaluated a duration-based baseline model using the same dynamic thresholding but operating on word durations instead of word scores where for λ = 0.25 the performance reached κ = 0.47. In all, the findings for C-PROM are similar to those of CGN, indicating that integration of probabilities over word durations leads to substantially better performance than integration over fixed-length windows.

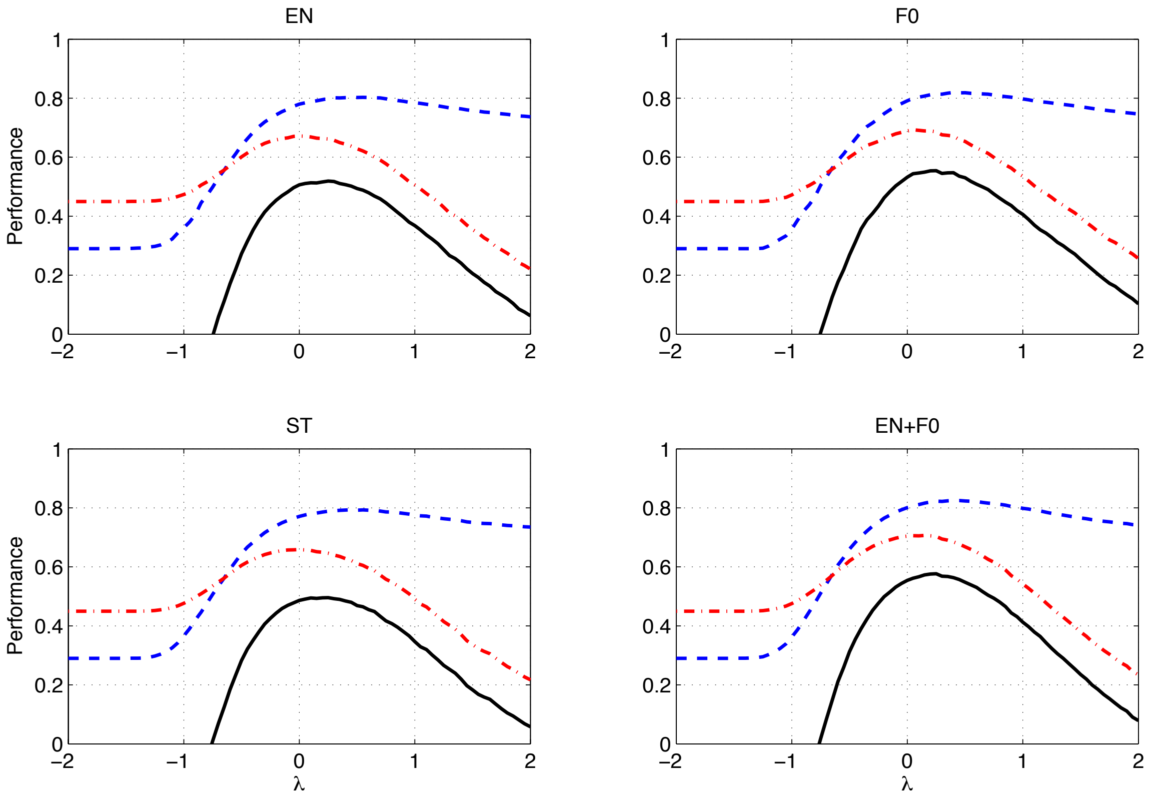

Fig. 9. Performance based on probability integration over word durations for individual features and their best combination for the C-PROM reference, for $Q = 16$ and $n = 2$, and as a function of the detection threshold λ. Blue dashed line represents accuracy, red dash-dotted line represents F-score, and black solid line represents the Fleiss kappa agreement.

Table 5. Performance of the word-level integration tests for $n = 1$, 2, and 3, $Q = 16$, $\lambda = 0.25$. Values in bold indicate the best results for each measure.

Features	κ	ACC	PRC	RCL	F
***n* = 1**					
EN	0.50	0.80	0.65	0.64	0.65
F0	0.39	0.77	0.64	0.47	0.54
ST	0.47	0.79	0.63	0.61	0.62
EN+F0	0.52	0.81	0.68	0.64	0.65
EN+ST	0.53	0.81	0.66	0.67	0.67
ST+F0	0.51	0.80	0.66	0.64	0.65
EN+F0+ST	0.54	0.81	0.68	0.68	0.68
***n* = 2**					
EN	0.52	0.80	0.65	0.68	0.66
F0	0.55	0.81	0.67	0.70	0.69
ST	0.50	0.79	0.63	0.66	0.65
EN+F0	**0.58**	**0.82**	**0.69**	**0.72**	**0.71**
EN+ST	0.54	0.81	0.65	0.70	0.68
ST+F0	0.55	0.81	0.66	0.71	0.68
EN+F0+ST	0.56	**0.82**	0.67	**0.72**	0.70
***n* = 3**					
EN	0.52	0.80	0.65	0.68	0.67
F0	0.56	**0.82**	0.67	0.70	0.69
ST	0.50	0.79	0.63	0.67	0.65
EN+F0	0.57	**0.82**	0.68	**0.72**	0.70
EN+ST	0.54	0.80	0.65	0.70	0.68
ST+F0	0.54	0.81	0.66	0.70	0.68
EN+F0+ST	0.56	0.81	0.67	**0.72**	0.69

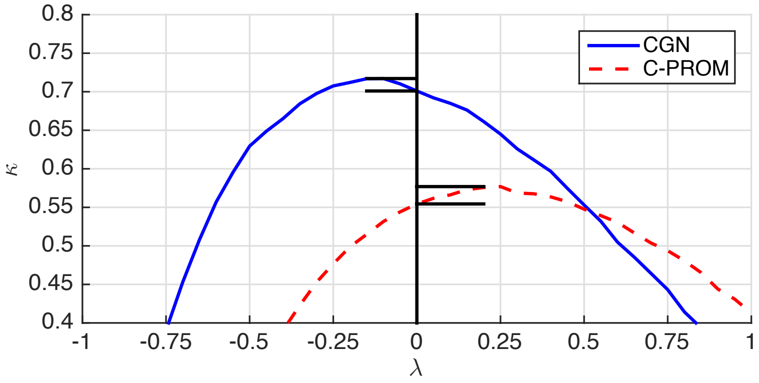

Fig. 10. Comparison of the optimal lambda values for the CGN (blue solid line) and C-PROM (red dashed line) corpora. X-axis: detection threshold λ. Y-axis: Fleiss-kappa performance on both corpora. Horizontal bars denote the optimal performance (higher) and performance at $\lambda = 0$ (lower) that is a compromise between the two corpora.

Since the optimal λ is somewhat different for the Dutch and French corpora (Tables 4 and 5), it was of interest whether a proper λ value is critical to the system performance. To this end, Fig. 10 shows the performance for both corpora as a function of λ (EN+F0, n = 2, Q = 16), separately indicating the threshold $\lambda = 0$ that provides a reasonable compromise between the two corpora. In case of $\lambda = 0$, kappa for CGN is 0.70 and for C-PROM it is 0.55, corresponding to a κ drop of 0.02–0.03 from the corpus-specific optimal value for λ. Although statistically significant, this difference is not qualitatively very large when considering the overall performance level of the system.

The difference in the optimal threshold likely reflects the difference in prominence distributions in the two corpora: CGN is very dense with prominent words, 42.6% of the words being marked as prominent. In contrast, in C-PROM, only 29% of all words are prominent. Since the detection threshold is dynamically determined according to the other words in the same utterance, a more stringent criterion (larger λ) is expected to lead to a smaller number of hypotheses per utterance, thereby fitting better to the C-PROM data. It appears that the current detection mechanism is not intelligent enough to account for such underlying distributional differences. Still, the performance is not greatly dependent on the exact value of the threshold, suggesting that the 3PRO is applicable to new languages without further tuning of λ. Although beyond the scope of the present study, further validation of the method with additional languages and/or speaking styles would help to determine the optimal cross-linguistic value for the threshold.

4.3. Comparison to supervised baselines

In order to compare the 3PRO performance to a situation where manual labeling of prominence is available, results for the supervised kNN, SVMs and CRFs were also computed for the same task. Table 6 shows the results for both CGN (top) and C-PROM (bottom) for all three systems with the best performing 3PRO results (word-level decoding) also shown as a reference.

Table 6. Performance of the supervised systems for CGN (top half) and C-PROM (bottom half). As a reference, the corresponding measures for the feature combination of EN+F0 with $Q = 16$, $n = 2$, $\lambda = -0.15$ for CGN and $\lambda = 0.25$ for C-PROM, are shown for the unsupervised word-level decoded 3PRO.

CGN	**kNN**	**SVM**	**CRF**	**3PRO EN+F0**
PRC	0.87	0.85	**0.86**	0.80
RCL	0.79	**0.87**	**0.87**	0.86
F	0.82	**0.86**	**0.86**	0.84
κ	0.70	0.75	**0.76**	0.72
ACC	0.85	**0.88**	**0.88**	0.86
C-PROM	**kNN**	**SVM**	**CRF**	**3PRO EN+F0**
PRC	0.78	**0.78**	0.77	0.69
RCL	0.60	**0.71**	0.70	0.72
F	0.67	**0.74**	0.73	0.71
κ	0.55	**0.63**	0.62	0.58
ACC	0.82	**0.84**	**0.84**	0.82

The supervised results are in line with previous findings on supervised prominence detection. For instance, Christodoulides and Avanzi (2014) report F-score of 78.4% for SVMs and F = 84.9% for a fusion of conditional random fields and random forests in prominence detection on French speech from the PFC corpus that contains both read and conversational speech (Durand et al., 2009). Similarly, Tamburini et al. (2014) report the best F-score of 77.0% for Italian using conditional neural fields, whereas Moubayed et al. (2010) report ACC = 72.55% for Swedish using SVMs in a ternary prominence classification task (*no*/*maybe*/*yes* manual labeling for prominence). To further ensure that our supervised baselines are valid, we evaluated C-PROM performance using the same division to training and test data as is described in Rosenberg et al. (2012), replicating their findings (Rosenberg et al. report ACC = 86.11% using similar features and L2-regularized logistic regression while our present CRF system achieved 85.73% on the same data using the CGN-optimized L2-regularization parameter).

In general, the 3PRO system compares well against the supervised systems even though it does not have access to prominence markings at any stage of the processing. The fully word-agnostic version achieves $\kappa = 0.64$ on CGN which is a reasonably high value considering the lack of the highly relevant word duration information (c.f., Table 2). When word boundaries are available during prominence decoding, 3PRO outperforms the kNN on both CGN and C-PROM with $\kappa = 0.72$ and $\kappa = 0.58$, respectively, and performs only slightly worse than the results obtained with the SVMs ($\kappa = 0.75$ / $\kappa = 0.63$) and CRFs ($\kappa = 0.75$ / $\kappa = 0.62$). All differences between the supervised baselines and the corresponding 3PRO performances are highly significant (χ^2 (1, $N_{\text{CGN}} = 7438$, $N_{\text{CPROM}} = 13184$) > 28 and $p << 0.001$ for all comparisons with McNemar's paired chi-square test).

5. Discussion and conclusions

Prominence has been widely studied with respect to its acoustic correlates that make its production and perception possible (see, e.g., Fry, 1955, 1958; Lieberman, 1960; Zhang, Nissen, & Francis, 2008). In addition, the connection between predictability of linguistic units in speech and perceptual prominence of these units has been documented earlier (see, e.g., Pan & Hirschberg, 2000). In the present work, we proposed an algorithm called 3PRO for the unsupervised detection of sentence prominence from speech, making use of the idea of the connection between prominence and predictability at the level of acoustic prosodic features. The algorithm is based on the idea that the short-term unexpected acoustic prosodic trajectories in speech will draw the listeners' attention and will therefore be perceived as prominent. This transforms the traditional approach of detecting certain value combinations of prominence-related acoustic cues to modeling of the probabilities of these cues given certain a priori experience with the language, suggesting that a system for prominence detection can be learned from data without access to prominence labels. Our current findings seem to support this assumption, with 3PRO performance reaching Fleiss kappa agreement of 0.64 with accuracy of 82.3% for the Dutch data and kappa of 0.49 with accuracy of 77.5% for the French data for the purely unsupervised system. When word boundaries are known during detection, the algorithm reaches kappa of 0.72 and accuracy of 85.5% for the Dutch data and kappa of 0.58 with accuracy of 82.3% for the French data, both measured with respect to human perception of sentence prominence in the same data. The result for both the Dutch and French data is superior to the classification performance of a supervised kNN-based system and close to SVMs and CRFs on the same data when the hyperparameters of the supervised systems are optimized for maximal performance on the prominence detection task.

The results also suggest that the proposed algorithm offers comparable or even improved performance over existing unsupervised approaches in prominence detection. For instance, in the study of Kalinli and Narayanan (2007), 78.1% accuracy is reported at the word level whereas Wang and Narayanan (2007) report 80% precision also at the word level (see also Tamburini, 2003). Moreover, 3PRO seems to also compare well with results obtained from other studies in supervised prominence detection. For instance, Rosenberg et al. (2015) report an accuracy of 89.03% using BiRNNs, Sridhar et al. (2008) report 86% accuracy using combinations of acoustic and syntactic features, and Wang and Narayanan (2007) report precision of 82.1% using SVMs, all studies evaluating prominence at the word level. Additionally, there is a study from Streefkerk et al. (1997) on the Dutch Polyphone corpus using artificial neural networks (ANNs) on a prominence detection task, reporting 82.1% accuracy at the word level. In general, these results suggest that 3PRO can be useful in, e.g., under-resourced languages where labeled training data for supervised systems do not exist, providing a principled way to detect words that stand out from their context in terms of their prosodic characteristics.

However, direct comparison of the results is difficult due to the differences in evaluation metrics and in the corpora and languages used in the experiments. An obvious source of disparity between studies with different datasets is the style and consistency of prominence annotations utilized on the data, with naïve listeners showing notably different prominence transcription patterns from trained annotators (see, e.g., Mo, Cole, & Lee, 2008; Breen, Dilley, Kraemer, & Gibson, 2012). There is no commonly agreed standard corpus for prominence evaluation, and access to the potentially most widely used Boston University Radio Speech Corpus (Ostendorf, Price, & Shattuck-Hufnagel, 1995) is greatly limited by the high price of the corpus. In addition, our present experiments show that the performance measures are highly dependent on the question of how the annotations from multiple independent transcribers are used as a reference. Here the best agreement between the current algorithm and the reference was obtained when it was sufficient for only one of the two annotators to mark a word as a prominent target (the BA reference). In contrast, the mean agreement with respect to each annotator separately is notably lower (the TH reference). In general, the less the annotators agree with each other, the larger the difference between the TH and BA metrics can potentially become. However, no commonly agreed method for creating a single reference from multiple annotations exists.

An additional practical aspect of the proposed method is the flexibility over the choice of integration frames. If needed, the method can operate in a purely unsupervised manner, however, the decoding stage can be also constrained by using information regarding the underlying linguistic units, leading to improved performance. The importance of linguistic grounding, at least at the level of words, became evident in the second experiment where acoustic prosodic expectations were integrated over word durations. In this case, prominence agreement increased from the purely unsupervised kappa of 0.64 to 0.72 for Dutch and from 0.49 to 0.58 for French, confirming that word duration is an important cue for the perception of prominence (see also the study of Moubayed et al., 2010).

In terms of the performance of the individual features, energy and F0 were the strongest cues for prominence, with their combination producing the best performance in both languages ($\kappa_{\mathrm{EN+F0}}$ =0.64 and 0.72 for CGN and 0.58 for C-PROM). In the case of spectral tilt, the overall contribution in predicting prominence was low in experiment 1 ($\kappa_{\mathrm{ST,CGN}}$=0.33, $\kappa_{\mathrm{ST,C\text{-}PROM}}$=0.32) whereas in experiment 2 its contribution was higher ($\kappa_{\mathrm{ST,CGN}}$=0.60, $\kappa_{\mathrm{ST,C\text{-}PROM}}$=0.50) due to the inherent inclusion of durational cues during the integration process, but still not showing complementary information with respect to F0 and energy. Although spectral tilt is suggested to be a correlate of prominence in Dutch (see, e.g., the study of Sluijter & van Heuven), we could not verify its contribution in the current experimental setup. In a similar manner, the study of Streefkerk, Pols, and ten Bosch (1999) on the Dutch Polyphone corpus did not find spectral tilt as a highly predictive feature for prominence. In all, the most predictive features for prominence in the current study seem to be duration, energy, and F0, a finding that is in accordance with the existing literature on prominence in other languages (see, e.g., Cutler, 2005; Werner & Keller, 1994; Fry, 1955, 1958).

Another interesting aspect in the present study was the temporal scale of the integration window used in the non-linguistically grounded decoding. It was found that the best average performance (Fig. 5 and 7) in both Dutch and French was obtained using window length of approximately 200 ms. Temporal integration windows of 200 ms (similar also to typical syllable duration) are also suggested in the literature to be the way how perceptual information is analyzed (see, e.g., Zwislocki, 1960; see also Sussman, Winkler, Kreuzer, Saher, Näätänen, & Ritter, 2002). Thus, this finding may also suggest a connection between how acoustic prosodic information is analyzed in temporal chunks in the absence of any type of linguistic or paralinguistic information. However, based on the current data, it is not possible to make any further inference beyond position this statement as a point for further investigation.

One potential limitation of the present work is that it focuses on local short-term dependencies at the acoustic feature-level. Recent studies using Latent-Dynamic CRFs (LDCRFs) have shown that hidden dynamics might exist in the sequences of non-prominent and prominent syllables (see, e.g., Tamburini et al., 2014; Cutugno et al., 2012), indicating that there might be a recurring rhythmic grouping of the prominent units across time. This is also discussed in the work of Arvaniti (2009) where it is argued that the rhythmic categorization of all languages should be based on one universal principle that could rest on the grouping and patterns of prominence (see also Dilley and McAuley, 2008, for an example; see also Fraisse, 1982). Future work should therefore extend 3PRO towards this direction. For instance, rhythmic regularities in prominence could be used to create another level of anticipatory organization that would be also modeled in a probabilistic manner, thus predicting the temporal positions of upcoming prominent words or syllables given the preceding detection outcomes. This information could be then used to modulate the detection threshold in a manner analogous to having prior probabilities for prominence on syllabic or word units.

In all, 3PRO is a computationally simple and therefore easy-to-use algorithm for prominence detection, operating in a manner similar to how prominence perception in humans is hypothesized to take place (see Kakouros & Räsänen, in press). The results indicate that it can achieve high agreement with annotator's markings, reaching a performance level close to what supervised approaches on the same data can attain. Furthermore, the findings suggest that the low-level (acoustic) short-term expectations appear to be one useful cue to prominence and therefore a potentially suitable first approximation in purely unsupervised approaches. As our current findings are based on two corpora of Dutch and French speech, more languages should be tested in future in order to verify that the algorithm generalizes well to other languages and speaking styles.

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 project "Computational modeling of language acquisition". An early version of part of this work was published in the proceedings of the Interspeech-2015 conference, Dresden, Germany, under the title "Automatic Detection of Sentence Prominence in Speech Using Predictability of Word-Level Acoustic Features".

References

Ananthakrishnan, S., & Narayanan, S. (2006). Combining acoustic, lexical, and syntactic evidence for automatic unsupervised prosody labeling. *In Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech-2006), Pittsburgh, Pennsylvania*, pp. 297–300.

Ananthakrishnan, S., & Narayanan, S. (2007). Improved speech recognition using acoustic and lexical correlates of pitch accent in a n-best rescoring framework. *In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-2007)*, *Honolulu*, *Hawaii*, pp. 873–876.

Ananthakrishnan, S., & Narayanan S. (2008). Automatic prosodic event detection using acoustic, lexical, and syntactic evidence. *IEEE Transactions on Audio, Speech, and Language Processing,* 16(1), pp. 216–228.

Andreeva, B., Barry, W., & Koreman, J. (2014). A Cross-language Corpus for Studying the Phonetics and Phonology of Prominence. *In Proceedings of the International Conference on Language Resources and Evaluation (LREC-2014), Reykjavik, Iceland,* pp. 326–330.

Arnold, D., & Wagner, P. (2008). The influence of top-down expectations on the perception of syllable prominence. *In Proceedings of the 2nd ISCA Workshop on Experimental Linguistics (ExLing-2008)*, *Athens, Greece*, pp. 25–28.

Arnold, D., Wagner, P., & Baayen, R. H. (2013). Using generalized additive models and random forests to model German prosodic prominence. *In Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech-2013)*, *Lyon, France*, pp. 272–276.

Arvaniti, A. (2009). Rhythm, timing and the timing of rhythm. *Phonetica*, 66(1-2), pp. 46–63.

Avanzi, M., Simon, A. C., Goldman, J. P., & Auchlin, A. (2010). C-PROM: An Annotated Corpus for French Prominence Study. *In Proceedings of Speech Prosody (SP-2010) Workshop on Prosodic Prominence, Chicago, Illinois.*

Avanzi, M., Goldman, J. P., Lacheret-Dujour, A., Simon, A. C., & Auchlin, A. (2007). Méthodologie et algorithmes pour la détection automatique des syllabes proéminentes dans les corpus de français parlé. *Cahiers of French Language Studies*, 13(2), pp. 2–30.

Aylett, M. P., & Bull, M. (1998). The automatic marking of prominence in spontaneous speech using duration and part of speech information. *In Proceedings of the International Conference on Spoken Language Processing* (*ICSLP-98*), *Sydney, Australia*, pp. 2123–2026.

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

Baker, R. E., & Bradlow, A. R. (2009). Variability in word duration as a function of probability, speech style, and prosody. *Language and Speech*, 52(4), pp. 391–413.

Baumann, S. (2014). The importance of tonal cues for untrained listeners in judging prominence. *In Proceedings of the 10th International Seminar on Speech Production (ISSP-2014)*, *Cologne, Germany*, pp. 21–24.

Bell, A., Brenier, J. M., Gregory, M., Girand, C., & Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language*, 60(1), pp. 92–111.

Bock, J. K., & Mazzella, J. R. (1983). Intonational marking of given and new information: Some consequences for comprehension. *Memory & Cognition*, 11(1), pp. 64–76.

Breen, M., Dilley, L. C., Kraemer, J., & Gibson, E. (2012). Inter-transcriber reliability for two systems of prosodic annotation: ToBI (Tones and Break Indices) and RaP (Rhythm and Pitch). *Corpus Linguistics and Linguistic Theory 8*, 2, pp. 277–312.

Bruce, N. D., & Tsotsos, J. K. (2009). Saliency, attention, and visual search: An information theoretic approach. *Journal of Vision*, 9(3), pp. 1–24.

Buhmann, J., Caspers, J., van Heuven, V. J., Hoekstra, H., Martens, J. P., & Swerts, M. (2002). Annotation of prominent words, prosodic boundaries and segmental lengthening by non-expert transcribers in the Spoken Dutch Corpus. *In Proceedings of the International Conference on Language Resources and Evaluation (LREC-2002), Las Palmas, Spain*, pp. 779–785.

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

Calhoun, S. (2010a). How does informativeness affect prosodic prominence? *Language and Cognitive Processes*, 25(7–9), pp. 1099–1140.

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

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

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

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

Chen, K., Hasegawa-Johnson, M., Cohen, A., Borys, S., Kim, S. S., Cole, J., & Choi, J. Y. (2006). Prosody dependent speech recognition on radio news corpus of American English. *IEEE Transactions on Audio, Speech, and Language Processing,* 14(1), pp. 232–245.

Chen, M., Liu, W., Yang, Z., & Hu, P. (2012). Automatic Prosodic Events Detection Using a Two-Stage SVM/CRF Sequence Classifier with Acoustic Features. In Liu, C. L., Zhang, C., and Wang, L. (Eds.), *Proceedings of the Chinese Conference on Pattern Recognition*, Springer Berlin Heidelberg, pp. 572–578.

Christodoulides, G., & Avanzi, M. (2014). An Evaluation of Machine Learning Methods for Prominence Detection in French. *In Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech-2014), Singapore*, pp. 116–119.

Cutler, A. (2005). Lexical stress. In Pisoni, D. B., and Remez, R. E. (Eds.), *The handbook of speech perception*, Blackwell Publishing Ltd, pp. 264–289. Doi: 10.1002/9780470757024.ch11

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

Cutler, A. (1987). Speaking for listening. In Allport, A., MacKay, D. G., Prinz, W., Scheerer, E. (Eds.), *Language Perception and Production: Relationships between Listening, Speaking, Reading and Writing*, Academic Press, London, pp. 23–40.

Cutugno, F., Leone, E., Ludusan, B., & Origlia, A. (2012). Investigating Syllabic Prominence With Conditional Random Fields and Latent-Dynamic Conditional Random Fields. *In Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech-2012), Portland, Oregon,* pp. 2402–2405.

Dilley, L. C., & McAuley, J. D. (2008). Distal prosodic context affects word segmentation and lexical processing. *Journal of Memory and Language*, 59(3), pp. 294–311.

Duchateau, J., & Ceyssens, T. (2004). Use and evaluation of prosodic annotations in Dutch. *Proceedings of the International Conference on Language Resources and Evaluation (LREC-2004), Lisbon, Portugal*, pp. 1517–1520.

Durand, J., Laks, B., & Lyche, C. (2009). Le projet PFC: une source de données primaires structurées. In J. Durand, B. Laks et C. Lyche (Eds.): Phonologie, variation et accents du français. Paris: Hermès. pp. 19–61.

Endress, A. D., & Hauser, M. D. (2010). Word segmentation with universal prosodic cues. *Cognitive Psychology*, 61(2), pp. 177–199.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76, pp. 378–382.

Fraisse, P. (1982). Rhythm and tempo. In Deutsch, D. (Ed.), *The psychology of music*, Academic Press, New York, pp. 149–180.

Fry, D. B. (1958). Experiments in the perception of stress. *Language and Speech*, 1(2), pp. 126–152.

Fry, D. B. (1955). Duration and intensity as physical correlates of linguistic stress. *Journal of the Acoustical Society of America*, 27(4), pp. 765–768.

Guinaudeau, C., & Hirschberg, J. (2011). Accounting for prosodic information to improve ASR-based topic tracking for TV broadcast news. *In Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech-2011), Florence, Italy*, pp. 1401–1404.

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

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

Jaeger, T. F. (2006). *Redundancy and Syntactic Reduction in Spontaneous Speech*. PhD thesis, Stanford University, Stanford, CA.

Jaeger, T. F., & Levy, R. P. (2007). Speakers optimize information density through syntactic reduction. In Schölkopf, B., Platt, J. C., and Hoffman, T. (Eds.), *Advances in neural information processing systems 19*, MIT Press: Cambridge, MA, pp. 849–856.

Jeon, J. H., & Liu, Y. (2012). Automatic prosodic event detection using a novel labeling and selection method in co-training. *Speech Communication*, 54(3), pp. 445–458.

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

Jurafsky, D. (1996). A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science*, 20(2), pp. 137–194.

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

Kakouros, S., & Räsänen, O. (2014b). Perception of Sentence Stress in English Infant Directed Speech. *In Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech-2014)*, *Singapore*, pp. 1821–1825.

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

Kakouros, S., & Räsänen, O. (2016). Statistical Learning of Prosodic Patterns and Reversal of Perceptual Cues for Sentence Prominence. *In Proceedings of the 38th Annual Conference of the Cognitive Science Society*, *Philadelphia, Pennsylvania*.

Kakouros, S., Räsänen, O., & Laine, U. K. (2013). Attention based temporal filtering of sensory signals for data redundancy reduction. *In Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP-2013)*, *Vancouver, Canada*, pp. 3188–3192.

Kakouros, S., & Räsänen, O. (in press). Perception of sentence stress in speech correlates with the temporal unpredictability of prosodic features. *Cognitive Science*. doi: 10.1111/cogs.12306

Kalinli, O., & Narayanan, S. (2007). A saliency-based auditory attention model with applications to unsupervised prominent syllable detection in speech. *In Proceedings of the Annual Conference of the International Speech Communication Association (Interspeeh-2007)*, *Antwerp, Belgium*, pp. 1941–1944.

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

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

Kohler, K. J. (2008). The perception of prominence patterns. *Phonetica*, 65(4), pp. 257–269.

Koreman, J., Andreeva, B., Barry, W. J., Sikveland, R. O., van Dommelen, W. (2009). Cross-language differences in the production of phrasal prominence in Norwegian and German. *In Proceedings of the Xth Conference of Nordic Prosody*, *Helsinki, Finland*, pp. 139–150.

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

Lam, T. Q., & Watson, D. G. (2010). Repetition is easy: Why repeated referents have reduced prominence. *Memory & Cognition*, 38(8), pp. 1137–1146.

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

Larson, M., & Jones, G. J. F. (2011). Spoken content retrieval: A survey of techniques and technologies. *Foundations and Trends in Information Retrieval,* 5(4-5), pp. 235–422.

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

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

Linde, Y., Buzo, A., & Gray, R. M. (1980). An algorithm for vector quantizer design. *IEEE Transactions on Communications,* 28(1), pp. 84–95.

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

Maier, A. K., Hönig, F., Zeißler, V., Batliner, A., Körner, E., Yamanaka, N., & Nöth, E. (2009). A language-independent feature set for the automatic evaluation of prosody. *In Proceedings of the Annual Conference of the International Speech Communication Association (Interspeeh-2009)*, *Brighton, UK*, pp. 600–603.

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

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 Annual Conference of the International Speech Communication Association (Interspeech-2002), Denver, Colorado*, pp. 745–748.

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

Mo, Y., Cole, J., and Lee, E. K. (2008). Naïve listeners' prominence and boundary perception. *In Proceedings of Speech Prosody (SP-2008), Campina, Brazil,* pp. 735–738.

Moniz, H., Mata, A. I., Hirschberg, J., Batista, F., Rosenberg, A., & Trancoso, I. (2014). Extending AuToBI to prominence detection in European Portuguese. *In Proceedings of Speech Prosody (SP-2014), Dublin, Ireland.*

Moubayed, A. S., Ananthakrishnan, G., & Enflo, L. (2010). Automatic prominence classification in swedish. *In Proceedings of Speech Prosody 2010 (SP-2010), Workshop on Prosodic Prominence, Chicago, Illinois*.

Obin, N., Lacheret-Dujour, A., & Rodet, X. (2008). French Prominence: a Probabilistic Framework. *In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-2008), Las Vegas, Nevada*, pp. 3993–3996.

Obin, N., Rodet, X., & Lacheret-Dujour, A. (2009). A Syllable-Based Prominence Detection Model Based on Discriminant Analysis and Context-Dependency. *In Proceedings of the International Conference on Speech and Computer (SPECOM-2009), St-Petersbourg, Russia.*

Oostdijk, N. H. J., Goedertier, W., Van Eynde, F., Boves, L., Martens, J.-P., Moortgat, M., and Baayen, H. (2002). Experiences from the Spoken Dutch Corpus project. *In Proceedings of the Third International Conference on Language Resources and Evaluation (LREC-2002)*, *Las Palmas, Gran Canaria*, pp. 340–347.

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

Ostendorf, M., Price, P. J., & Shattuck-Hufnagel, S. (1995). The Boston University radio news corpus. *Technical Report ECS-95-001*, Boston University.

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

Patil, S., Arsikere, H., & Deshmukh, O. (2015). Acoustic Stress Detection for Improved Navigation of Educational Videos. *In Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech-2015)*, *Dresden, Germany*, pp. 1882–1883.

Racca, D. N., & Jones, G. J. F. (2015). Incorporating Prosodic Prominence Evidence into Term Weights for Spoken Content Retrieval. *In Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech-2015)*, *Dresden, Germany*, pp. 1378–1382.

Rosenberg, A. (2010). AutoBI-a tool for automatic toBI annotation. *In Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech-2010), Makuhari, Chiba, Japan*, pp. 146–149.

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

Rosenberg, A., Fernandez, R., & Ramabhadran, B. (2015). Modeling Phrasing and Prominence Using Deep Recurrent Learning. *In Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech-2015), Dresden, Germany*, pp. 3066–3070.

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

Räsänen, O., Doyle, G., & Frank, M. C. (2015). Unsupervised word discovery from speech using automatic segmentation into syllable-like units. *In Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech-2015), Dresden, Germany*, pp. 3204–3208.

Shattuck-Hufnagel, S., & Turk, A. E. (1996). A prosody tutorial for investigators of auditory sentence processing. *Journal of psycholinguistic research*, 25(2), pp. 193–247.

Silverman, K. E. A., Beckman, M. E., Pitrelli, J. F., Ostendorf, M., Wightman, C. W., Price, P., Pierrehumbert, J. B., & Hirschberg. J. (1992). TOBI: a standard for labeling English prosody. *In Proceedings of the Second International Conference on Spoken Language Processing (ICSLP-1992). Banff, Alberta, Canada,* pp. 867–879.

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

Smith, C. (2011). Perception of prominence and boundaries by naïve French listeners. *In Proceedings of the 17th International Congress of Phonetic Scences (ICPhS-2011)*, *Hong Kong, China*, pp. 1874–1877.

Streefkerk, B. M., Pols, L. C., & Ten Bosch, L. F. (1997). Prominence in read aloud sentences, as marked by listeners and classified automatically. *In Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam*, vol. 21, pp. 101–116.

Streefkerk, B. M., Pols, L. C., & ten Bosch, L. (1999). Acoustical features as predictors for prominence in read aloud Dutch sentences used in ANN's. *In Proceedings of the European Conference* on *Speech* Communication and Technology (*EUROSPEECH-1999), Budapest, Hungary*, pp. 551–554.

Sussman, E., Winkler, I., Kreuzer, J., Saher, M., Näätänen, R., & Ritter, W. (2002). Temporal integration: intentional sound discrimination does not modulate stimulus-driven processes in auditory event synthesis. *Clinical Neurophysiology*, 113(12), pp. 1909–1920.

Sridhar, V. K. R., Bangalore, S., & Narayanan, S. S. (2008). Exploiting acoustic and syntactic features for automatic prosody labeling in a maximum entropy framework. *IEEE Transactions on Audio, Speech, and Language Processing,* 16(4), pp. 797–811.

Szaszák, G., Beke, A., Olaszy, G., & Tóth, B. P. (2015). Using Automatic Stress Extraction from Audio for Improved Prosody Modelling in Speech Synthesis. *In Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech-2015), Dresden, Germany*, pp. 2227–2231.

Tamburini, F., Bertini, C., & Bertinetto, P. M. (2014). Prosodic prominence detection in Italian continuous speech using probabilistic graphical models. *In Proceedings of Speech Prosody (SP-2014), Dublin, Ireland*, pp. 285–289.

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

Tamburini, F. (2003). Automatic prosodic prominence detection in speech using acoustic features: an unsupervised system. *In Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech-2003)*, *Geneva, Switzerland*, pp. 129–132.

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

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

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

Van Son, R. J., & Pols, L. C. (2003a). How efficient is speech? *In Proceedings of the Institute of Phonetic Sciences, University of Amsterdam*, 25, pp. 171–184.

Van Son, R. J., & Pols, L. C. (2003b). Information structure and efficiency in speech production. *In Proceedings of the Annual Conference of the International Speech Communication Association* (*Interspeech-2003)*, *Geneva, Switzerland*, pp. 769–772.

Wagner, P., Tamburini, F., & Windmann, A. (2012). Objective, Subjective and Linguistic Roads to Perceptual Prominence. How are they compared and why? *In Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech-2012)*, *Portland, Oregon,* pp. 2394–2397.

Wagner, P., Origlia, A., Avesani, C., Christodoulides, G., Cutugno, F., D'Imperio, M., ... & Moniz, H. (2015). Different parts of the same elephant: a roadmap to disentangle and connect different perspectives on prosodic prominence. *In Proceedings of the International Congress of Phonetic Sciences (ICPhS-2015), Glasgow, Scotland.*

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

Watson, D. G., Arnold, J. E., & Tanenhaus, M. K. (2008). Tic Tac TOE: Effects of predictability and importance on acoustic prominence in language production. *Cognition*, 106(3), pp. 1548–1557.

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

Wightman, C. W., & Ostendorf, M. (1994). Automatic labeling of prosodic patterns. *IEEE Transactions on Speech and Audio Processing,* 2(4), pp. 469–481.

Yoon, T., Chavarria, S., Cole, J., & Hasegawa-Johnson, M. (2004). Intertranscriber reliability of prosodic labeling on telephone conversation using toBI. *In Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech-2004), Jeju Island, Korea*, pp. 2729–2732.

You, H. J. (2012). Determining prominence and prosodic boundaries in Korean by non-expert rapid prosody transcription. *In Proceedings of Speech Prosody (SP-2012)*, *Shanghai, China*, pp. 318–321.

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

Zwislocki, J. (1960). Theory of temporal auditory summation. *The Journal of the Acoustical Society of America*, 32(8), pp. 1046–1060.

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