Comparison of spectral tilt measures for sentence prominence in speech – effects of dimensionality and adverse noise conditions

Sofoklis Kakouros, Okko Räsänen & Paavo Alku

Department of Signal Processing and Acoustics, Aalto University, P.O. Box 12200, 00076 AALTO, Finland

Contact information:

Sofoklis Kakouros (corresponding author). Email: sofoklis.kakouros@aalto.fi. Department of Signal Processing and Acoustics, Aalto University, P.O. Box 12200, FI-00076, AALTO, Finland.

Okko Räsänen, okko.rasanen@aalto.fi. Department of Signal Processing and Acoustics, Aalto University, P.O. Box 12200, FI-00076, AALTO, Finland.

Paavo Alku, paavo.alku@aalto.fi. Department of Signal Processing and Acoustics, Aalto University, P.O. Box 12200, FI-00076, AALTO, Finland.

Abstract

Linguistic prominence in speech is known to correlate with the acoustic measures of energy, F0, and duration. In contrast, the role of spectral tilt in the realization of prominence has remained more inconsistent between previous empirical investigations. This may be partially due to the lack of a standard method for quantifying spectral tilt or due to difficulties in estimating the acoustical source of spectral tilt, the glottal flow, from continuous speech. These issues have rendered interpretations and comparisons between studies difficult. In addition, (i) little is known about the robustness of tilt estimators for prominence detection in the case when speech is not clean but corrupted, as in real life, by environmental noise or telephone transmission (i.e. degradation caused by bandpass filtering and quantization noise). Moreover, (ii) little attention has been paid to multidimensional representations of source spectrum that can potentially incorporate more information about the phonation style than purely scalar measures. In this work, we study spectral tilt in signaling prominence in spoken Dutch and French under different levels of additive noise, and for telephone-band coded speech, and compare several one-dimensional tilt measures that have been previously encountered in the literature as well as multidimensional tilt measures. We also compare spectral tilt measures with other standard acoustic correlates for prominence, namely, energy, F0, and duration. Our results provide further empirical support for the finding that tilt is a systematic correlate of prominence in Dutch, that the role is smaller in French, and that energy, F0, and duration appear still to be the most robust features for discriminating prominent and non-prominent words. In addition, our results show that there are notable differences between different tilt measures at different levels of noise, and that multidimensional representations for tilt improve class separability from the scalar measures.

Keywords: prosody, sentence prominence, acoustic measures, spectral tilt, noise robustness, DNN

1. Introduction

Spoken language contains a multitude of distinct information types at different levels, ranging from linguistic content to speaker information, that are all intertwined into a single representational form, the acoustic speech signal. One defining difference between spoken and written language is that spoken language carries prosodic information. Prosody involves the use of suprasegmental properties that allow for a broader way of expression that goes beyond the coding of information into individual phonemic or lexical items, and conveys information at slower rates extending across individual segments. In this regard, prosody delivers information on *how* something is spoken as opposed to *what* is spoken (or written), and therefore prosody contributes substantially to the naturalness of speech. In addition, prosody impacts the intelligibility of speech, as prosodic cues are used to guide perceptual parsing of the speech stream. In this context, *prominence* is a prosodic phenomenon that can be generally defined as the property by which a linguistic unit is perceived to be standing out from its environment (see, e.g., Cutler, 2005; Terken & Hermes, 2000; Wagner et al., 2015; Shattuck-Hufnagel & Turk, 1996, for related definitions). As the definition for prominence is specific to the linguistic domain upon it is evaluated, a particular description for *sentence prominence* encompasses the degree of perceived emphasis for one or more words during a sentence (see, e.g., Cole, Mo, & Hasegawa-Johnson, 2010) and for *lexical prominence* the accentuation of syllables within words (see, e.g., Cutler, 2005). It is also important to note that there are several terminological variants in the literature to denote the phenomenon, such as stress and emphasis, to name a few (see also Wagner et al., 2015, for a discussion). In this work, we will use the term *prominence* to refer to the perceptual impression of standing out, as also defined in the work of Terken and Hermes (2000).

Prominence is an important constituent of speech serving several functions in discourse and speech perception. It is therefore a particularly important component for natural language applications (see, e.g., Mehrabani, Mishra, & Conkie, 2013; Racca & Jones, 2015). For instance, prominence can convey information about the pragmatic context of the discourse, reflecting the speaker's intent to mark specific words as the targets of information focus (Bolinger, 1972). The most widely acknowledged function of prominence across studies is in signaling the information status of a word (see, e.g., Calhoun, 2010; Cole et al., 2010; Wagner & Watson, 2010). This means that prominent words often introduce information that is *new* or important in the discourse. On the other hand, words that lack prominence are seen as *given*, referring to information that can be accessed situationally or anaphorically (see, e.g., Brown, 1983), that is, referring to information that is immediately accessible, through, for instance, the context of the preceding discourse. Beyond the communicative role, other studies have investigated the linguistic function and the impact of prominence during sentence comprehension (see, e.g., Bock & Mazzella, 1983; Cutler & Foss, 1977; Terken & Nooteboom, 1987). The general finding from these studies is that prominence facilitates speech comprehension through faster processing of the prominent targets (e.g., Bock & Mazzella, 1983; Terken & Nooteboom, 1987). In addition, prominence can be indicative of factors such as the lexical class of words in a sentence (see, Wagner et al., 2015, for a discussion).

Earlier research on prominence has identified four acoustic features that are correlated with the production and perception of prominent units in speech: signal energy (e.g., Fry, 1955; Kochanski, Grabe, Coleman, & Rosner, 2005; Lieberman, 1960), fundamental frequency (F0) (e.g., Fry, 1958; Lieberman, 1960; Terken, 1991), duration (e.g., Fant & Kruckenberg, 1994; Fry, 1955; Lieberman, 1960), and spectral tilt (e.g., Campbell & Beckman, 1997; Heldner, 2001; Sluijter & van Heuven, 1996a, 1996b). For instance, already in the early works of Fry (1955, 1958) and Lieberman (1960), it was found that variations in duration (e.g., longer syllable duration) are important for prominence, with increased unit duration correlating with increased prominence of the unit. The relation between prominent units and F0 seems to be more complex, as simply the magnitude of F0 change or distance of F0 maxima to the baseline do not seem to sufficiently describe this relationship (Terken, 1991; see also Kakouros & Räsänen, 2016; Gussenhoven, Repp, Rietveld, Rump, & Terken, 1997; Rietveld & Gussenhoven, 1985). Moreover, there seems to be a competition between duration and F0 in conveying the impression of prominence to the listener (see, e.g., Niebuhr & Winkler, 2017). Nonetheless, there is strong correlational evidence of the importance of F0 in conveying prominence in speech (see, e.g., Kohler, 2008). As for the role of signal amplitude, Kochanski et al. (2005) have shown a strong independent role of energy in predicting prominence with some studies also indicating trading relationships of the feature with duration (see, e.g., Fry, 1955, 1958; Gay, 1978; Turk & Sawusch, 1996). The acoustic operationalization of loudness on the basis of signal energy has been considered as a limiting factor in some studies, as it cannot account for the energy allocation across frequency bands that is known to affect the perceived loudness of auditory input (see, e.g., Sluijter, van Heuven, & Pacilly, 1997). This might potentially limit the capability of the scalar energy measure to characterize the relevant prosodic spectral cues used by listeners. Therefore, spectral tilt has been utilized as a measure to reflect the differences between the higher and lower frequency bands, and some studies have found it to provide cues for the discrimination between prominent and non-prominent units in speech (see, e.g., Sluijter & van Heuven, 1996a). However, not all studies have been able to empirically validate the contribution of spectral tilt in the task (see, e.g., Campbell & Beckman, 1997; Kochanski et al., 2005), and the contribution of energy, F0, and duration to the prominence phenomenon seems to be better established than that of spectral tilt. This might be explained by the fact that languages (and hence studies) simply differ in their use of these cues for conveying prominence. However, it might also be the case that there are established standard ways to measure energy, duration, and F0 (albeit to a lesser degree) than what is available for spectral tilt, and therefore the tilt measures between studies are not always directly comparable. In addition, some measures of tilt are directly computed from the acoustic speech signal while others attempt to quantify spectral balance of the glottal excitation, the former being potentially confounded by the linguistic content of speech.

1.1 Spectral tilt and prominence

A variety of methods have been proposed in the literature to measure spectral tilt that are also often encountered under different but closely related terms (e.g., spectral balance, spectral tilt, spectral emphasis). In addition, there seems to be no consensus on connecting specific measures to the utilized terminology or the underlying phenomenon being measured. For instance, some studies may use the term spectral tilt with reference to the spectral slope of the *excitation* of the human speech production mechanism (i.e. the glottal volume velocity waveform generated by the vocal folds) while others might use it with reference to the spectral slope of the system's *output* (i.e. the speech pressure signal where the source, tract, and lip radiation are coupled). In this work we will use the term *source tilt (SOT)* in order to denote the slope of the voice source spectrum and *surface tilt (SUT)* to denote the slope of the combined spectrum of the voice source, vocal tract, and lip radiation. Moreover, there are several methods to measure tilt that follow different overall procedures resulting into measures that are not exactly equivalent but that all attempt to quantify the superficially same phenomenon (the relative contribution of high versus low frequency bands of the spectrum). As a result, these measures are not necessarily directly comparable, and the implications of the potential differences among the tilt estimators are currently largely unknown, especially in typical real-world listening and recording conditions where speech is also corrupted with various types of additive and channel noise.

Overall, the diversity of the measures quantifying spectral tilt poses important challenges in the interpretation of results across different studies. For instance, several studies have investigated the utilization of measures for SUT. Specifically, Sluijter and van Heuven (1996a) measured spectral tilt as the band-limited intensity difference across four continuous spectral bands (0–0.5, 0.5–1, 1–2, and 2–4 kHz). In another study, Campbell and Beckman (1997) used the harmonic ratio (difference in dB between the first and second harmonic of F0, H1-H2) in order to quantify a measure for spectral tilt. Other studies use an array of different methods, including calculation of the difference in dB between the overall intensity and the intensity of the fundamental frequency (or of the intensity in a frequency band centered at F0) (Barbosa, Eriksson, & Åkesson, 2013; Eriksson, Thunberg, & Traunmüller, 2001; Heldner, 2001), taking the first cepstral coefficient (C1) (Tsiakoulis, Potamianos, & Dimitriadis, 2010), taking the difference in dB between a signal with high-frequency pre-emphasis and flat frequency weighting (SPLH-SPL) (Fant, Kruckenberg, Liljencrants, & Hertegård, 2000), taking the difference in dB between the first harmonic and third formant (H1-F3) (Okobi, 2006), fitting a regression line in the magnitude spectrum (Aronov & Schweitzer, 2016; Lu & Cooke, 2009), taking the band-limited spectral energy ratios (Murphy, McGuigan, Walsh, & Colreavy, 2008; Prieto & Ortega-Llebaria, 2006), using the long-term average spectrum (LTAS) to obtain band-limited energy ratios (Sundberg & Nordenberg, 2006), and using all-pole modeling techniques (Magi, Pohjalainen, Bäckström, & Alku, 2009).

In addition, some studies utilize similar measures, such as regression line fitting and harmonic ratio, but, instead of applying the measures directly on the short-term spectrum of speech (such as in the case of SUT), they utilize the spectrum of the glottal source waveform obtained through glottal inverse filtering (GIF) (see, e.g., Iseli et al., 2006; Jackson, Ladefoged, Huffman, & Antoñanzas-Barroso, 1985; Kreiman, Gerratt, & Antoñanzas-Barroso, 2007). Other studies make use of various parameterizations of the voice source, such as the Liljencrants-Fant (LF) model (Fant, Liljencrants, & Lin, 1985), in order to derive a measure for tilt (see, e.g., Fant & Kruckenberg, 1994) and may also use other parameters of the voice source in order to study and evaluate different prosodic phenomena (see, e.g., Fant & Kruckenberg, 1994; Iseli et al., 2006). In general, it is important to note that the contribution of the voice source in the task of discriminating prominence categories, and prosody in general, has remained largely undetermined. This is largely due to the fact that the glottal source waveform is hard to quantify as it is not directly observable, therefore rendering the accurate estimation of the voice source signal challenging. In addition, due to this inherent limitation to reliably and automatically estimate the glottal source waveform, many studies have earlier relied on labor-intensive methods that required manual optimization (see, Kane & Gobl, 2013, for a discussion) making the analysis on large volumes of data problematic. However, current technology also enables automatic voice source estimation from the speech signal using GIF techniques such as the Quasi-closed phase analysis (QCP) (Airaksinen, Raitio, Story, & Alku, 2014), and an increasing number of studies have utilized this possibility for investigating prosodic phenomena. For instance, Ní Chasaide, Yanushevskaya, Kane, and Gobl (2013) proposed a so-called voice prominence hypothesis (VPH), suggesting that prominence lending accented syllables may be dependent on a number of source parameters (including F0). Their results support VPH, indicating a connection between changes in the voice source parameters and changes in the degree of accentuation (Ní Chasaide et al., 2013; see also Yanushevskaya, Gobl, Kane, & Ní Chasaide, 2010; Yanushevskaya, Murphy, Gobl, & Ní Chasaide, 2016).

An additional factor that renders investigations of the voice source challenging is that the estimation of the glottal volume velocity waveform is very sensitive to noise, whereas the typically used measures, such as the F0, appear to be overall more robust[1]. It is therefore of interest to make use of noise-robust methods for the estimation of the glottal volume velocity waveform as this also reflects the majority of real-life situations where speech is typically produced in the presence of different types of noise. For instance, a deep neural network (DNN) -based system for robust spectral tilt estimation was recently described by Jokinen and Alku (2017) that enables the estimation of the glottal source tilt in non-ideal signal conditions without explicitly performing GIF in the estimation phase. Instead, the relationship between the speech power spectrum and the underlying glottal excitation signal is learned in a supervised manner by the neural network from training data consisting of glottal flow signals estimated by GIF from clean speech. In this way, the DNN enables the estimation of the glottal volume velocity waveform through non-linear statistical regression from the speech spectrum.

As there exists no single default method to compute spectral tilt but rather many approaches, with differences observed at several levels, the goal of the present work is to gain a better understanding of the performance differences of a distinct set of tilt measures in characterizing prominence. This study builds upon an earlier effort (Kakouros, Räsänen, & Alku, 2017) that investigated differences between tilt measures, which, however, was limited in scope, and focused only on one corpus and select scalar measures. In this work, the aim is: (i) to compare the most well-known measures for spectral tilt together with a newly-proposed DNN-based technique (Jokinen & Alku, 2017) for prominence classification in speech, (ii) to evaluate whether multidimensional tilt representations bring performance improvements in the task, (iii) to compare the relative importance of the evaluated tilt measures with respect to the widely acknowledged acoustic correlates of prominence of energy, F0, and duration, and (iv) to examine the performance of tilt measures under non-ideal conditions met when processing signals in noisy real-life scenarios. The study is conducted using clean and corrupted speech in two languages (Dutch and French) by involving two types of corruption (additive noise and telephone band coding).

[1] Note that many GIF techniques require F0 and/or glottal closure instant estimates as a part of their operation, and are therefore inherently limited in their performance by the reliability of the F0 estimation.

2. Data

In this work, a total of three different speech corpora were utilized. Specifically, two corpora, CGN (Dutch) and C-PROM (French), consisting of continuous speech from two phonologically distinct languages, were used as the basis for all evaluations, as they include prosodic annotations for prominence. In addition, a third corpus, the Phonetic Corpus of Estonian Spontaneous Speech (Lippus, Tuisk, Salveste, & Teras, 2013), was used for the purpose of providing high-quality speech recordings needed for the method described in section 3. All three corpora are described in the next subsections in more detail.

2.1. CGN

The Spoken Dutch Corpus (Corpus Gesproken Nederlands; CGN) is a corpus of contemporary standard Dutch as spoken by adults in the Netherlands and Flanders, containing nearly 9 million words (800 hours of speech). The corpus includes manually generated or verified annotations such as phonetic transcriptions, word level alignment, and prosodic annotations (see Duchateau, Ceyssens, & van Hamme, 2004; Oostdijk et al., 2002, for a more detailed description). For the present evaluations, the Dutch news broadcast (*component-k*) part of the corpus was utilized, consisting of 5088 news broadcasts (≈27.4 hours of speech data) and spoken by 29 speakers (22 male and 7 female). *Component-k* includes a prosodically annotated subset consisting of 134 news broadcasts spoken by 10 different speakers (9 male and 1 female) (≈44.3 minutes of speech data). The annotations were hand-labeled using binary (prominent/non-prominent) markings by two trained annotators (see Buhmann et al., 2002, for more details), containing a total of 7438 word tokens.

2.2. C-PROM

Sentence prominence was also studied from continuous speech in French by using the C-PROM corpus (Avanzi, Simon, Goldman, & Auchlin, 2010) that is specifically annotated for prominence studies. The corpus 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 comprises 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 their labeling, resulting in a single set of prominence markings for the data (see Avanzi, Goldman, Lacheret-Dujour, Simon, & Auchlin, 2007, for more details).

2.3. EstPhon

In order to provide a source for high-quality clean speech training signals for DNN-based estimation of glottal volume velocity waveforms (section 3), we utilized the Phonetic Corpus of Estonian Spontaneous Speech of the University of Tartu[2] (*EstPhon*, see Lippus et al., 2013, for more details). The database consists of high-quality recordings of Estonian spontaneous speech between conversing test subjects and was recorded using near-field microphones. The corpus comprises different types of phonetic segmentations including, for instance, manually verified syllable annotations. The corpus contains a total of 60 hours of recordings by speakers from different age groups, dialectological, and social backgrounds. In this work, we used 1165 randomly chosen utterances from the studio section of the corpus for the DNN training.

3. Methods

3.1. Estimation of acoustic features

3.1.1 Energy, F0, and word duration

Energy, F0, and word duration were used as the reference features in this work, as it has been well established across a number of studies that they correlate well with the manifestation of prominence in speech (see, e.g., Fry, 1955, 1958; Kochanski et al., 2005; Kohler, 2008; Lieberman, 1960; Terken, 1991). In order to compute them, speech data were initially downsampled to 16 kHz. F0 estimation was carried out using a noise robust pitch tracker (Drugman & Alwan, 2011) with a 100-ms window and 10-ms hop size. The pitch tracker provided pitch estimates as well as a voicing decision for each frame of the analysis. Energy was computed using a 20-ms window and 10-ms hop size, and word durations were extracted directly from the corpora annotations.

3.1.2 Spectral tilt measures

For the comparative analysis of the spectral tilt measures, a number of different tilt estimation techniques that are commonly encountered in the literature were utilized. In this work, beyond the standard scalar one-parameter models, we also include in the analysis four multidimensional

[2] Information about the Phonetic Corpus of Estonian Spontaneous Speech is also available at http://www.keel.ut.ee/en/languages-resourceslanguages-resources/phonetic-corpus-estonian-spontaneous-speech.

features. All tilt measures are described in more detail in Tables 1 and 2 and were computed over a 20-ms window and using a 10-ms hop size.

Table 1: Definitions of surface tilt (SUT) measures utilized in this study, where D denotes the dimensionality of the features.

Tilt measure	D	Definition
H1-H2 (dB)	1	Difference in dB between the first and second harmonic (see, e.g., Campbell & Beckman, 1997).
H1-F3 (dB)	1	Difference in dB between the first harmonic and third formant (see, e.g., Okobi, 2006).
C1	1	The first Mel-frequency cepstral coefficient (MFCC; see, e.g., Tsiakoulis, Potamianos, & Dimitriadis, 2010).
SER	1	Spectral energy ratio (in dB) between 0–1 kHz and 1–5 kHz (see, e.g., Murphy et al., 2008).
SLF	1	Slope of the line obtained by fitting a first order polynomial to the short-term logarithmic magnitude spectrum of speech (spectral regression – see, e.g., Aronov, & Schweitzer, 2016).
LP1	1	First order forward linear prediction coefficient (LPC).
SLF6D	6	Coefficients of a sixth order polynomial fitted to the short-term logarithmic magnitude spectrum of speech.

Table 2: Definitions of source tilt (SOT) measures utilized in this study, where D denotes the dimensionality of the features.

Tilt measure	D	Definition
QCP	1	Slope of the line fit to the logarithmic magnitude spectrum of the glottal volume velocity waveform obtained from quasi-closed phase glottal inverse filtering (Airaksinen et al., 2014).
DNNC	1	Slope of the line fit to the short-term logarithmic magnitude spectrum of the DNN-estimated glottal volume velocity waveform.
QCP6D	6	Coefficients of a sixth order polynomial fitted to the logarithmic short-term magnitude spectrum of the glottal volume velocity waveform obtained from quasi-closed phase glottal inverse filtering (Airaksinen et al., 2014).
DNNC6D	6	Coefficients of a sixth order polynomial fitted to the logarithmic magnitude spectrum of the DNN-estimated glottal volume velocity waveform.

3.1.3 DNN-based spectral tilt estimation

Jokinen and Alku (2017) recently proposed a method to estimate and parameterize the glottal source spectrum in noisy, non-ideal conditions where conventional GIF analysis cannot be used due to its known sensitivity to noise (Alku, 2011). This method used a feed-forward DNN to map an input feature vector (the logarithmic speech power spectrum) into an output vector (all-pole model of the glottal source spectrum parameterized using line spectrum frequencies, or LSFs). In this work, a DNN was trained for the prediction of the source LSFs, describing the glottal source spectrum directly from the logarithmic FFT magnitude spectrum of the speech input (20-ms window, 10-ms hop size; see also Fig. 1). The DNN was trained on high-quality clean speech only. In our earlier study, augmentation of the training data with noise-corrupted versions of the signals was also investigated (Kakouros, Räsänen, & Alku, 2017). Since this did not lead to performance improvements, only clean training is utilized in the current setup. It is worth emphasizing that the DNN-based spectral tilt estimation method does not require GIF in the estimation phase. GIF is used only in the training phase to estimate the glottal flow from studio-quality speech. The spectrum of the estimated glottal flow is then parameterized by the DNN using the LSF feature vector. In the current study, the quasi closed phase (QCP) method (Airaksinen et al., 2014) was used as the GIF algorithm in training of the DNN.

Before training, the 255-dimensional spectral frame inputs and 8-dimensional LSF outputs of the DNN were z-score normalized across all training data to ensure proper scaling. The implementation of the utilized feed-forward neural network consisted of a configuration layout of d = [64, 32, 16] hidden units per layer, sigmoid activation function for the hidden layers, a linear output layer, a learning rate of 0.1, 100 epochs, minibatch size of 1000, and mean squared error (MSE) as the cost function. This resulted in a single DNN for tilt prediction based on clean speech (DNNC). The final tilt estimates used in the comparisons were then obtained by fitting a first (DNNC) or sixth (DNNC6D) order polynomial in the spectrum of the glottal waveform as parametrized by the predicted LSFs. The order of the multidimensional DNN measure, and also of the multidimensional measures in Tables 1 and 2 (see also Jokinen and Alku, 2017, for a comparison with other multidimensional measures), was selected as a compromise between a low enough dimensionality to avoid detailed fitting to the formant structure yet high enough to comply with the order of the DNN-based SOT method described in Jokinen and Alku (2017) and utilized here. We also compare the resulting tilt estimates to those computed directly from speech using QCP.

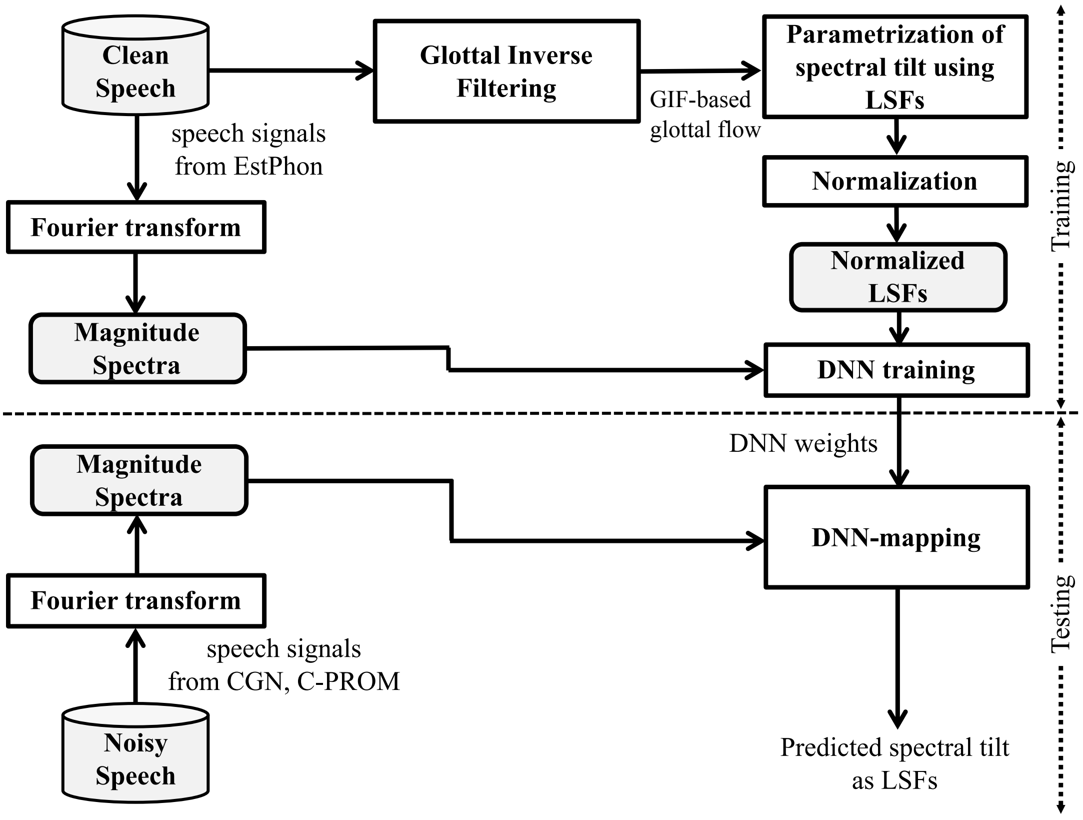

Figure 1: *Schematic diagram of the training/testing process for the DNN-based tilt estimator.*

3.2 Word-level statistical descriptors of the acoustic features

All evaluations in this study were carried out at the word level on the CGN and C-PROM corpora. Specifically, the manually labelled prominence markings were used to divide the data into two categories: prominent and non-prominent words. The original CGN data contain labels by two annotators, thus we considered all words with at least one prominence marking as prominent (see Kakouros & Räsänen, 2016, for a similar approach). For the C-PROM corpus, the original prominence labelling 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 that were used as the reference (see Avanzi, Goldman, Lacheret-Dujour, Simon, & Auchlin, 2007, for more details). As C-PROM contains syllable-level annotations, the syllable-level prominence labels were aligned with the word-level transcriptions in order to provide word-level binary prominence annotations (see also Kakouros & Räsänen, 2016; Rosenberg, Cooper, Levitan, & Hirschberg, 2012).

In order to evaluate and compare all the different acoustic features at the word level, five word-level statistical descriptors were computed for all but the duration feature: (i) mean, (ii) max, (iii) min, (iv) standard deviation (SD), and (v) range (max-min) of the given feature during the word (see, e.g., Christodoulides & Avanzi, 2014; Kakouros & Räsänen, 2016, for a similar approach). For the multidimensional features, these descriptors were computed independently for each of the feature dimensions.

3.3. Evaluation of prominent/non-prominent class separation

3.3.1 Separation of scalar measures for energy, F0, duration, and spectral tilt

In order to compare separability of the different features and their descriptors for the prominent and non-prominent classes, the estimated Z-score based effect-size *r* from Wilcoxon rank sum test (Eq. (1)) was utilized together with the symmetric Kullback–Leibler (KL) divergence (Eq. (2)). KL-divergence was computed by quantizing the data into *Q* = 25 discrete bins with a uniform number of samples in all bins across the entire data set. Both measures quantify the degree of separability of the prominent and non-prominent classes with zero corresponding to no difference. In Eq. (1) and Eq. (2), P_{pr} and P_{npr} denote the probability density of the matching bins and N_{pr} and N_{npr} denote the number of samples for the two classes, respectively.

$$r = \frac{Z}{\sqrt{N_{pr} + N_{npr}}} \tag{1}$$

$$D_{KL} = \sum P_{pr} \log(\frac{P_{pr}}{P_{npr}}) + \sum P_{npr} \log(\frac{P_{npr}}{P_{pr}}) \tag{2}$$

3.3.2 Separation of multidimensional measures and feature combinations

Two standard supervised classification methods, namely the *k*-nearest-neighbor (kNN) and support vector machines (SVMs), were used to compare multidimensional tilt features and feature combinations since statistical tests, such as the Wilcoxon rank sum test, can be applied only on populations with one-dimensional data. In addition, in order to obtain a benchmark of the overall performance in comparison to the class separation approach described in section 3.2.1, we also ran supervised classification for all scalar measures (where classification was run for a combination of all statistical descriptors, resulting in 5-dimensional vectors for the scalar measures). Both classifiers were trained and tested in an n-fold manner using the binary prominence labels and the corresponding word-level feature descriptors, and with the classification carried out in a context-independent manner for each word token. In particular, for CGN, supervised classification of words into prominent and non-prominent classes was run in a 10-fold classification procedure that was carried out by always training with data from 9 speakers and testing with a held-out talker (see section 2.1 for the corpus description). Correspondingly, for C-PROM, one recording was always used for testing while the remaining 23 recordings were used for training, resulting in a 24-fold evaluation procedure (see section 2.2 for the corpus description).

We wanted to ensure that the separability measures (classification accuracies) for different features were based on the same metrics while measuring the degree of class overlap under ideal (potentially non-linear) decision boundaries. Therefore, hyperparameters for kNN and SVMs were optimized for maximal average performance across all features on the test data, separately for each fold. Note that, in this case, overfitting of the hyperparameters is not a concern, as we simply want to measure class separability in our given sample, not to deploy a practical classifier for prominence. In practice, classification performance for kNN was computed for all values of *k* in the range between 1 and 20 and the average of the fold-specific best results (same *k* for all features of the same fold) is reported in the result section. For the SVM training, a radial basis kernel function was used. The SVM scaling factor σ and box-constraint C 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 $C = [0.001, 0.01, \ldots, 1000]$ to find an optimal combination of the two, using again the average performance across all features in the fold as the criterion.

For the evaluation of classification performance, precision (PRC), recall (RCL), their harmonic mean (F-value), and accuracy (ACC) were used as the main quality measures and were defined according to the equations below –where true positives (*tp*), true negatives (*tn*), false positives (*fp*), and false negatives (*fn*) define the classification outcomes:

$$\text{RCL} = tp \,/\, (tp + fn) \tag{3}$$

$$\text{PRC} = tp \,/\, (tp + fp) \tag{4}$$

$$\text{F} = (2 \text{ x PRC x RCL}) \,/\, (\text{PRC} + \text{RCL}) \tag{5}$$

$$\text{ACC} = (tp + tn) \,/\, (tp + fp + fn + tn) \tag{6}$$

In addition, Fleiss kappa (Fleiss, 1971) was used as a measure of the reliability of agreement between the human annotation reference for prominence and the classification output. The Fleiss kappa (FK) measure allows for comparison between the typical agreement for human annotators (see, e.g., Kakouros & Räsänen, 2014; Mo, Cole, & Lee, 2008). FK measures the degree of agreement between two or more annotators within a nominal scale of $\kappa \in [-1,1]$, where an outcome of $\kappa = 1$ indicates a full agreement and $\kappa \leq 0$ indicates no agreement ($\kappa = 0$ suggests that the number of agreements is what would be expected by chance).

3.4. Preparation of noisy and coded speech

For the purpose of the evaluations in non-ideal conditions, additive background noise and simulation of speech transmission in telephone networks (i.e. degradation caused by bandpass filtering and generation of quantization noise) were utilized. Specifically, noisy versions of the CGN and C-PROM signals were generated by corrupting them with additive babble noise (different signals from the ones used for corrupting the DNN training data) with SNRs of -10, -5, 0, 5, 10, 15, 20, 40, and 60 dB in addition to using clean speech. It is important to note here that the broadcast speech in CGN and the different discourse genres in C-PROM are inherently noisier than ideal studio-quality recordings (e.g., EstPhon data), and therefore "clean" refers to the (potentially non-ideal) signal quality where no further artificial degradations have been introduced.

In order to simulate signal degradation caused by telephone transmission, the original speech signals from both corpora were coded using the adaptive multi-rate (AMR) codec (see, European Telecommunications Standards Institute [ETSI], TS 126 090, TS 126 204, 2011, for more details). AMR is a speech compression method that was developed by the 3rd Generation Partnership Project (3GPP), standardized by ETSI, and that consists of several different codecs. AMR is widely used in digital cellular networks such as the Global System for Mobile Communications (GSM) and the Universal Mobile Telecommunications System (UMTS). In this work, we use the narrowband AMR codec (AMR-NB) that is used for transmission of the traditional telephone band speech in the range of 300–3400 Hz as well as the wideband version of the AMR (AMR-WB) that provides a bandwidth between 50–7000 Hz. Taken together, speech signals from both CGN and C-PROM were degraded using both AMR-NB and AMR-WB and tested in the experiments along with the additive noise conditions.

4. Experiments

The capability of the different tilt measures to discriminate between prominent and non-prominent words was evaluated in two experiments. The first experiment involved an evaluation over all individual one-dimensional (scalar) measures using class separability metrics, while the second consisted of an evaluation of all distinct multidimensional features and features combinations (with the addition of all scalar measures) using supervised classification of words into prominent and non-prominent classes. All reported results below are presented at the word level and separately for the CGN and C-PROM corpora.

4.1. Prominence class separation for scalar measures

After training the DNN using the Estonian corpus, the 134 annotated speech signals of CGN *component-k* and 24 annotated recordings of C-PROM corpus were used in order to compute all features from clean, noisy, and AMR-coded versions of CGN and C-PROM, respectively. Since the overall behavior of D_{KL} was found to be nearly identical to the effect size r across all conditions, only the latter is reported in the following sections.

4.1.1 Energy, F0, duration

4.1.1.1 Separability on CGN

The results for the standard acoustic features of energy, F0, and duration provide a strong indication of their importance in distinguishing between prominent and non-prominent categories in Dutch. An overview of the features’ performance for the five statistical descriptors is presented in Figure 2 for clean speech and different SNRs in the added noise condition and in Figure 3 for AMR-coded speech. A substantial variation in the overall class separation for the different descriptors can be observed, with *min* providing the least consistent and lowest performance across all evaluated statistical descriptors. Overall, it seems that measures of dispersion, such as the *SD* and *range*, enable a better characterization of the two prominence classes. In particular, it is possible that these descriptors (*range, SD*) can better capture the dynamic behavior of the features over words, also reflecting the inherent dynamic nature of the prosodic phenomena. On the other hand, *min* and *max* alone cannot capture dynamic properties of the features (over words) and might also lack in robustness as they are very sensitive to outliers –potentially also explaining the weak and variable separation that is observed for *min* in Figure 2. As for the *mean*, it provides the second weakest separation (see also Figure 3) and also has the highest reduction in performance for AMR-coded speech. On the whole, the best performing separation is achieved for the descriptors of *range* and *SD* with *max*, *mean*, and *min* following in order of relative performance.

For the individual features' performance, duration seems to be the most robust feature in characterizing prominence, reaching a class separation of r = 0.72. F0 and energy are also very important features for the task with, however, substantially lower performance. Specifically, for the *range* descriptor and clean speech, F0 reached $r = 0.56$ and energy $r = 0.46$. With decreasing SNR levels (Figure 2), all descriptors seem to be affected starting from 10 dB SNR and with the performance deteriorating from that point onwards. For AMR-coded speech (Figure 3), both F0 and energy are heavily impacted by the coding, with the performance significantly deteriorating from the baseline (note that the class separation for duration remains unimpaired as it is independent of the coding process).

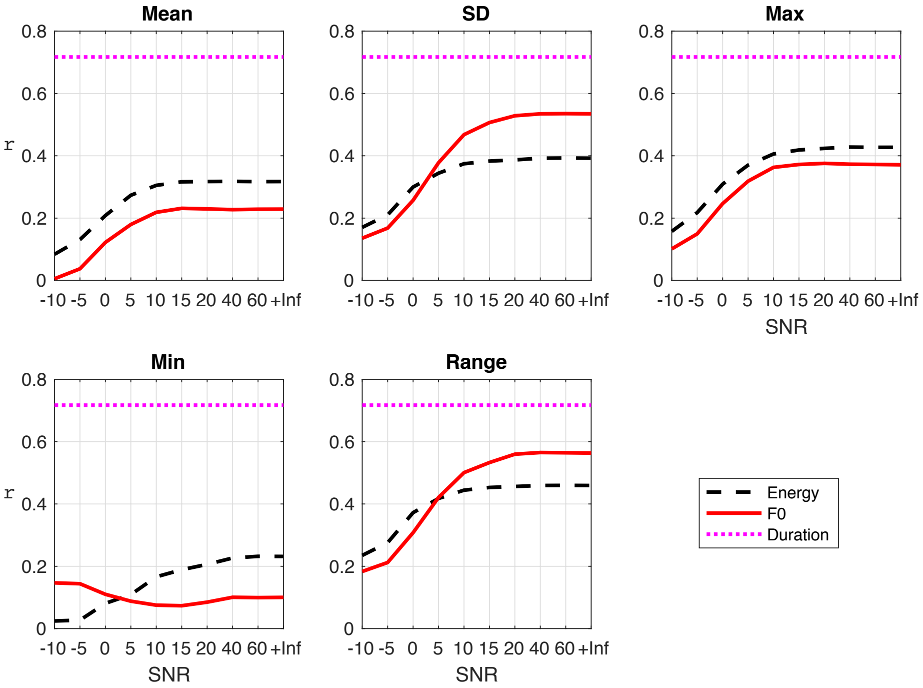

Figure 2: *Prominent and non-prominent class separation (r) for energy, F0, and duration plotted for mean, SD, max, min, and range. SNR varies from -10 dB to +Inf (clean signal). Data taken from CGN.*

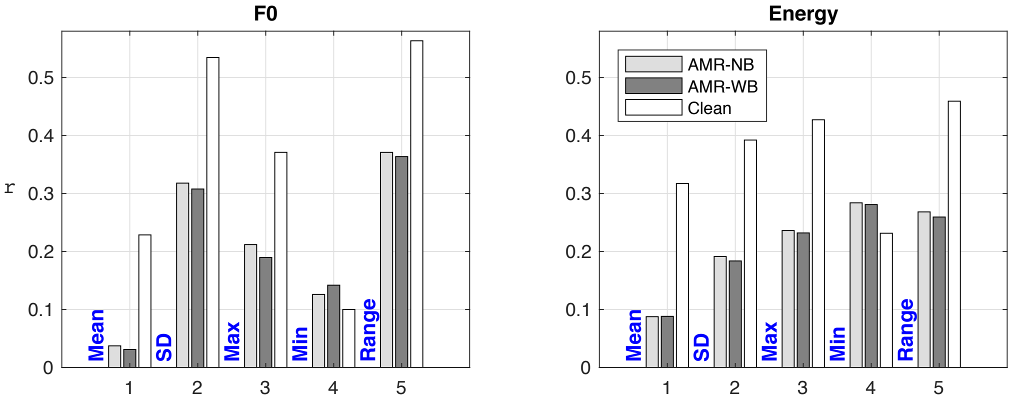

Figure 3: *Prominent and non-prominent class separation (r) for energy and F0 and for the mean, SD, max, min, and range. Data from CGN as clean, or coded using AMR-NB or AMR-WB.*

4.1.1.2 C-PROM

The results for the five statistical descriptors for C-PROM are similar to those for CGN. Specifically, the best performing statistical descriptors are the *range* and *SD* followed by *max*, *mean*, and *min* (in order of performance). An overview of the features' performance for all statistical descriptors is presented in Figure 4 for clean speech and different SNRs in the added-noise condition and in Figure 5 for AMR-coded speech. Similar to what was observed for CGN, the overall best performing feature is duration with a class separation of $r = 0.48$. For the remaining features, the best performance across all SNRs is attained for clean speech and the *range* descriptor with $r = 0.40$ for F0 and $r = 0.35$ for energy. Decreasing SNR affects separability from approximately 10 dB onwards. For AMR-coded speech, the results are somewhat surprising as the impact of the encoding process seems to be relatively limited on the class separations (opposite to what was observed for CGN). Specifically, for the *range* descriptor and for F0, AMR-NB attains $r = 0.38$, AMR-WB $r = 0.38$ whereas for clean speech the corresponding value was $r = 0.40$. For energy and the *range* descriptor, AMR-NB attains $r = 0.39$, AMR-WB $r = 0.39$, and for clean speech $r = 0.35$. This improvement in performance for energy (for AMR-coded speech), though unexpected, seems to be related with the processing steps involved in the AMR codec (possibly due to some type of gain scaling at the output), and also associated with the non-ideal signal quality of the "clean" C-PROM recordings. A closer inspection of the original C-PROM recordings revealed a sporadic incidence of signal clipping that was likely introduced by erroneously calibrated recording equipment and setup or by some post-processing step applied to the digitized signals before releasing the corpus. Examining the respective AMR-coded signals showed that the clipping distortion in both AMR-WB and AMR-NB was substantially reduced when compared to the original recordings. It seems that the processing steps in the AMR codec (e.g., gain scaling and gain smoothing; see Bessette et al., 2002, for more details) improved the overall quality of speech in C-PROM, at least at the signal level, and respectively allowed for a better prominence class separation performance for energy.

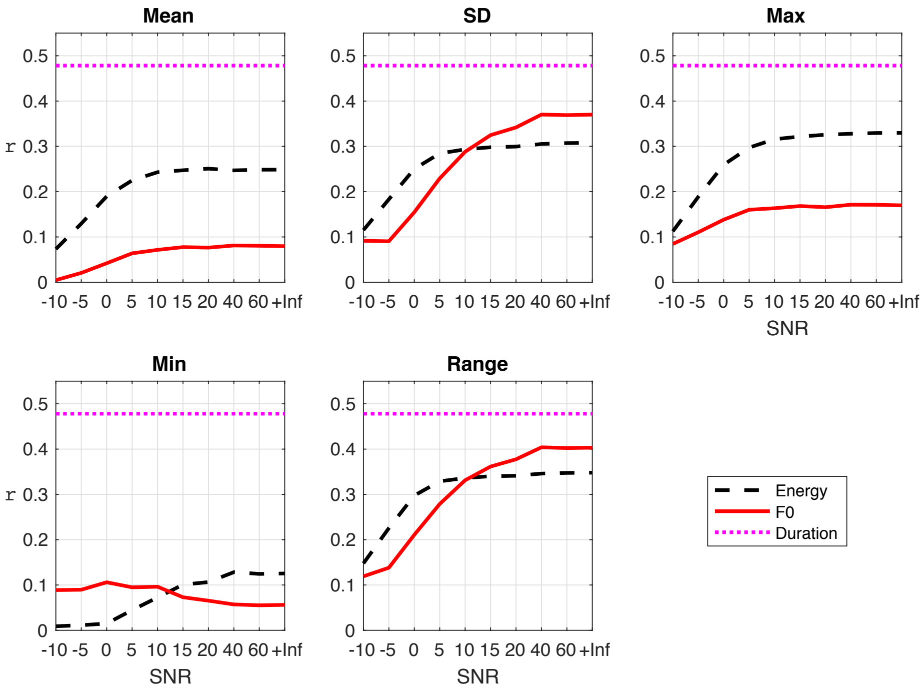

Figure 4: *Prominent and non-prominent class separation (r) for energy, F0, and duration plotted for mean, SD, max, min, and range. SNR varies from -10 dB to +Inf (clean signal). Data taken from C-PROM.*

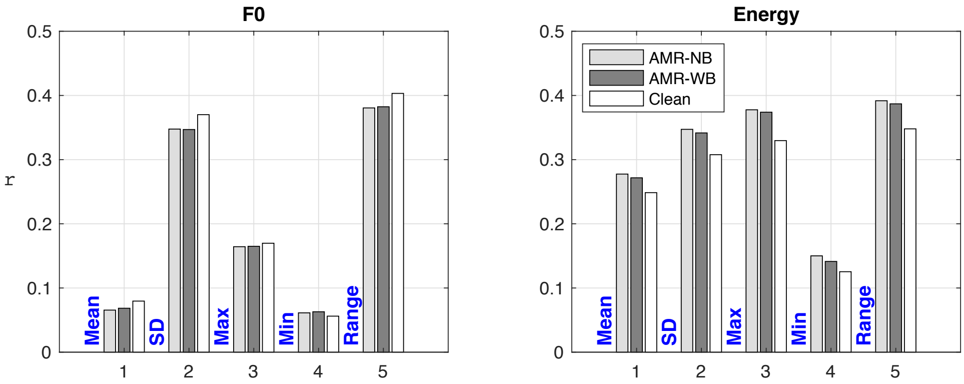

Figure 5: *Prominent and non-prominent class separation (r) for energy and F0 and for the mean, SD, max, min, and range. Data from C-PROM as clean, or coded using AMR-NB or AMR-WB.*

4.1.2 Separability of scalar tilt measures

A total of eight one-dimensional tilt measures (6 SUT and 2 SOT) were evaluated for CGN and C-PROM. Specifically, the class separation was evaluated for (i) H1-H2, (ii) H1-F3, (iii) LP1, (iv) C1, (v) SLF, (vi) SER, (vii) QCP, and (viii) DNNC at different SNR levels and for AMR-coded speech, as reported in the next subsections.

4.1.2.1 CGN

The results for different tilt measures on CGN can be seen in Figure 6. For clean speech, the best overall performance is attained for the *range* of C1 and DNNC with r = 0.44 and r = 0.45, respectively. Decreasing SNR levels start to have a major effect on class separation performance from approximately 10 dB SNR and reaching r = 0 at -5 dB. AMR coding of speech decreases the performance for both AMR-NB and AMR-WB for all tilt measures and across all statistical descriptors (see Table 3). For instance, separability of C1 *range* drops from r = 0.44 to r = 0.29 (AMR-NB) and to r = 0.30 (AMR-WB), and for DNNC *range* from r = 0.45 to r = 0.27 (AMR-NB) and to r = 0.28 (AMR-WB).

A closer look in the performance across the five statistical descriptors reveals a different behavior for tilt measures to what was observed for energy and F0 (where *range* and *SD* had the best overall performance). On the whole, the main finding is that *range* and *max* are the statistical descriptors that better capture the differences between the two prominence classes for tilt in CGN. *Range* is consistently the best descriptor among all tilt measures examined and *max* is the second. Variations in performance can also be observed across the descriptors where it is evident that not all tilt measures perform the same way across the different descriptors. For instance, for clean speech, H1-H2 and SER seem to be performing best for the *range* and *max* whereas for C1 and H1-F3 the respective best descriptors are the *range* and *min*. These variations are likely due to qualitative differences in the tilt measures reflected into their class separation performance.

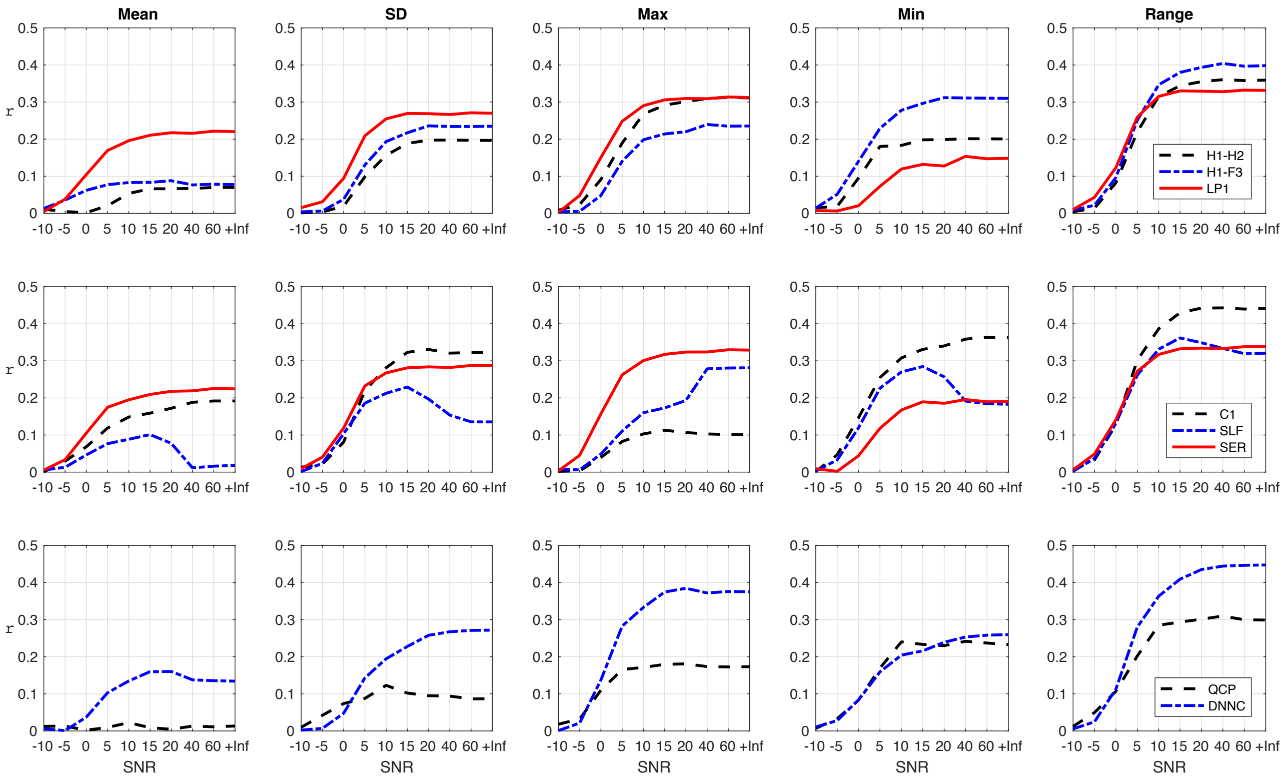

Figure 6: *Prominent and non-prominent class separation (r) for one-dimensional tilt measures plotted for mean, SD, max, min, and range. SNR varies from -10 dB to +Inf (clean signal). Data taken from CGN.*

Table 3: *Prominent and non-prominent class separation (r) for one-dimensional tilt measures for the mean, SD, max, min, and range. CGN data were coded using AMR-NB (NB) and AMR-WB (WB).*

	H1-H2		**H1-F3**		**LP1**		**C1**		**SLF**		**SER**		**QCP**		**DNNC**	
	NB	WB	NB	WB	NB	WB	NB	WB	NB	WB	NB	WB	NB	WB	NB	WB
Mean	0.01	0.01	0.00	0.00	0.08	0.08	0.00	0.01	0.02	0.02	0.09	0.09	0.01	0.00	0.03	0.02
SD	0.18	0.18	0.17	0.18	0.20	0.19	0.16	0.21	0.15	0.15	0.18	0.18	0.10	0.12	0.13	0.12
Max	0.18	0.19	0.18	0.20	0.21	0.20	0.19	0.18	0.19	0.19	0.20	0.20	0.19	0.21	0.21	0.22
Min	0.20	0.21	0.21	0.21	0.18	0.20	0.19	0.18	0.20	0.20	0.19	0.20	0.19	0.19	0.17	0.21
Range	**0.28**	**0.29**	**0.28**	**0.30**	**0.25**	**0.24**	**0.29**	**0.30**	**0.27**	**0.28**	**0.22**	**0.23**	**0.24**	**0.26**	**0.27**	**0.28**

4.1.2.2 C-PROM

Figure 7 presents an overview of the results for C-PROM. Among all scalar tilt measures evaluated and for clean speech, the best overall class separability was attained for the *range* of C1, H1-F3, and DNNC, with $r = 0.32$, 0.30, and 0.28, respectively (see Figure 8). Similar to CGN, decreasing SNR levels have a major effect on class separation performance from approximately 10 dB SNR, and all measures reach $r = 0$ at -5 dB. AMR coding of speech, however, does not seem to significantly impact the class separation performance for tilt measures. A similar observation was also noted for F0 earlier in section 4.1.1.2 for the same corpus. For instance, separability of C1 *range* drops from $r = 0.32$ in clean speech to $r = 0.28$ with AMR-NB and to $r = 0.30$ with AMR-WB, whereas for DNNC the separability drops from $r = 0.28$ (clean speech) to $r = 0.26$ (AMR-NB) and to $r = 0.28$ (AMR-WB). As in the case for F0 and energy earlier (see discussion in section 4.1.1.2), this difference may be due to the encoding process and non-ideal signal quality of the recordings. Finally, class separation performance across the statistical descriptors seems to behave similarly to that of CGN with the difference that, in addition to *range* and *max*, *SD* seems to be performing approximately at the same level with *max*. As in CGN, the same type of variation in performance for the different tilt measures across descriptors is also observed for C-PROM.

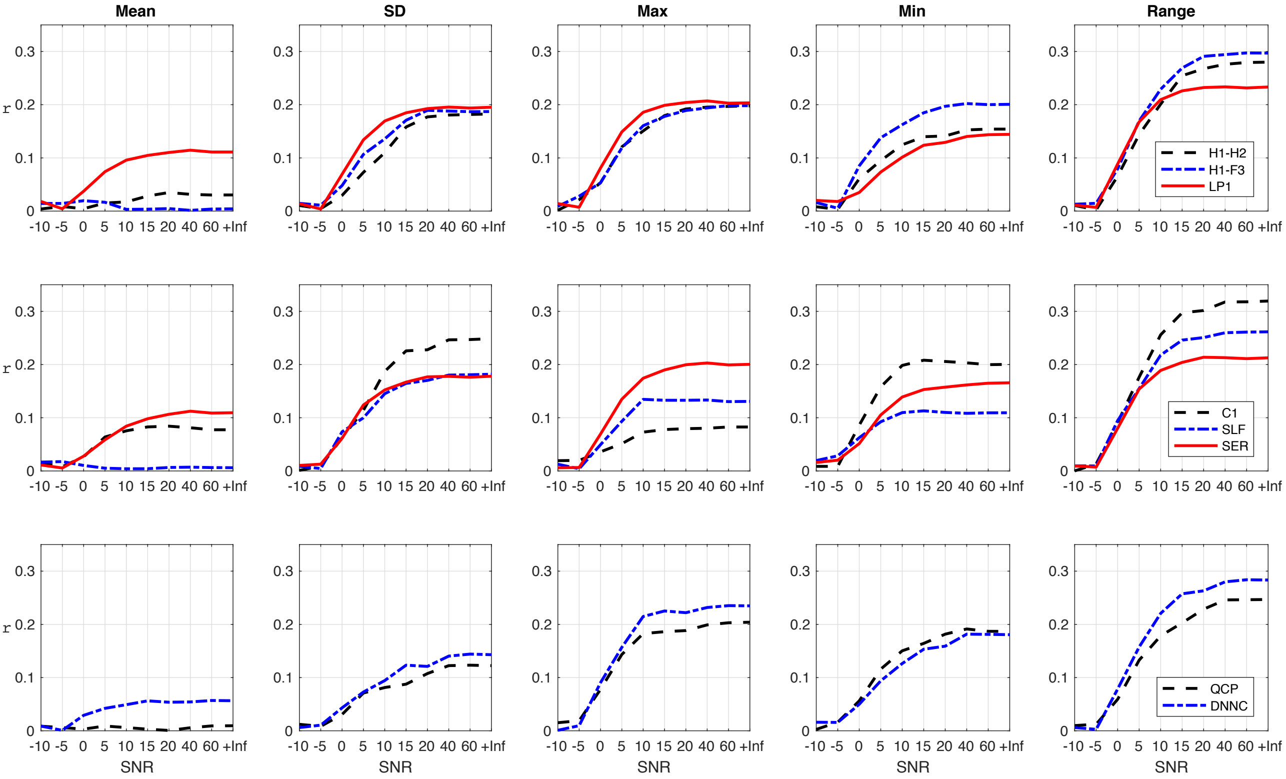

Figure 7: *Prominent and non-prominent class separation (r) for one-dimensional tilt measures plotted for mean, SD, max, min, and range. SNR varies from -10 dB to +Inf (clean signal). Data taken from C-PROM.*

Table 4: *Prominent and non-prominent class separation (r) for one-dimensional tilt measures for the mean, SD, max, min, and range. C-PROM data were coded using AMR-NB (NB) and AMR-WB (WB).*

	H1-H2		**H1-F3**		**LP1**		**C1**		**SLF**		**SER**		**QCP**		**DNNC**	
	NB	WB	NB	WB	NB	WB	NB	WB	NB	WB	NB	WB	NB	WB	NB	WB
Mean	0.02	0.02	0.00	0.01	0.12	0.11	0.09	0.06	0.04	0.01	0.12	0.11	0.01	0.01	0.01	0.03
SD	0.17	0.18	0.17	0.17	0.20	0.20	0.17	0.23	0.11	0.17	0.17	0.18	0.11	0.11	0.14	0.14
Max	0.16	0.16	0.16	0.17	0.22	0.21	0.11	0.12	0.20	0.17	0.20	0.21	0.19	0.19	0.20	0.23
Min	0.13	0.14	0.18	0.19	0.14	0.16	0.24	0.20	0.16	0.16	0.16	0.17	0.19	0.19	0.20	0.19
Range	**0.27**	**0.28**	**0.27**	**0.28**	**0.24**	**0.24**	**0.28**	**0.30**	**0.23**	**0.27**	**0.21**	**0.22**	**0.23**	**0.23**	**0.26**	**0.28**

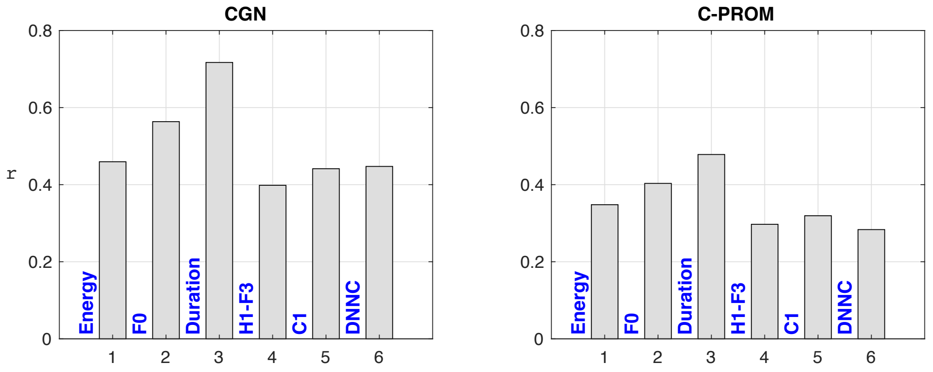

Figure 8: *Prominent and non-prominent class separation (r) for the best configuration of clean speech and range descriptor. Features included are energy, F0, duration, and the three best performing one-dimensional tilt measures. Data taken from CGN and C-PROM.*

4.2. Prominence classification performance using multidimensional measures

Two standard supervised classification methods (kNN and SVMs) were used for word prominence classification on CGN and C-PROM in order to understand the capability of the multidimensional tilt measures and feature/descriptor combinations to discriminate prominent and non-prominent words in speech. In addition, all one-dimensional measures were also evaluated in this experiment for comparison. In the experiments, for the one-dimensional measures, all five statistical descriptors were concatenated into a five-dimensional vector and, respectively, for the multidimensional (6D) measures the five statistical descriptors were concatenated into a 30-dimensional vector. The results are presented in the next subsections.

4.2.1 CGN

Overall, the performance of the multidimensional tilt measures appears to be substantially better than that of their one-dimensional counterparts. An overview of the results can be seen in Figures 9 and 10 for both SVM and kNN using the Fleiss kappa as the measure of agreement between classifier hypotheses and annotated ground truth. Contrary to the scalar measures, the multidimensional tilt measures seem to have more invariable performance across SNRs with the exception of *mean* which performs the poorest with agreement levels in the range of κ = 0 and 0.38 (not plotted here separately). As can be seen in Figure 9, the best overall performance is attained for *range* and *max* and for the measures of DNNC6D and SLF6D. In particular, for SVM, clean speech, and the *max* descriptor, DNNC6D reaches κ = 0.50 and SLF6D κ = 0.46.

Combining the five statistical descriptors adds robustness in the measures' behaviour across different SNRs and also further boosts their overall classification performance. For instance, DNNC6D reaches κ = 0.54 and SLF6D κ = 0.55 whereas the corresponding best performance for the one-dimensional measures for SVM (all descriptors combined) was DNNC κ = 0.50 and SLF κ = 0.47 –note that the best performance for a single descriptor (*range*, clean speech) was for DNNC κ = 0.41 and SLF κ = 0.27. As in the case of the multidimensional measures, inclusion of all descriptors for the one-dimensional measures also improves their performance (see Figure 10). However, even with the addition of all descriptors, their performance remains lower than that obtained for the multidimensional measures. For instance, the best performing one-dimensional tilt measures with combined descriptors are C1 and H1-H2, reaching κ = 0.50 and 0.48, respectively. Taken together, the performances of all multidimensional tilt measures are at the level of energy and F0 with κ = 0.55 and 0.56, respectively (with combined descriptors, see also Figure 10). Decreasing levels of SNR affect the performance of multidimensional measures similarly to what was observed for one-dimensional measures, with the highest degradation taking place below 10 dB SNR. Correspondingly, AMR coding of speech has a major impact on tilt measures with performance dropping to κ = 0.31, 0.33 for AMR-NB and κ = 0.32, 0.34 for AMR-WB, for the measures of DNNC6D and SLF6D, respectively.

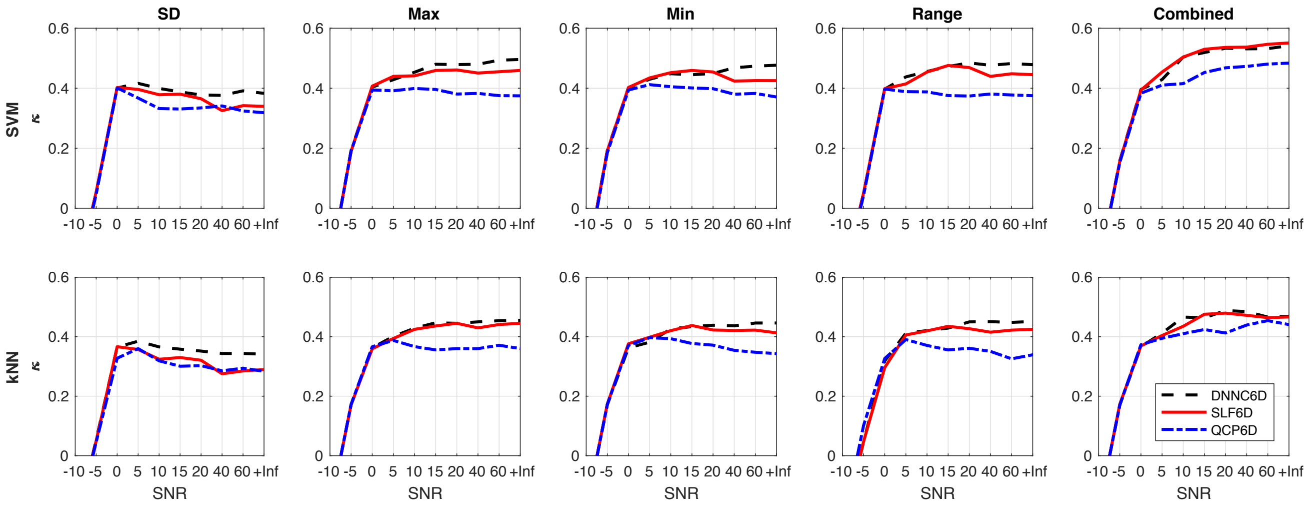

Figure 9: *SVM and kNN Fleiss kappa (κ) performance for multidimensional tilt measures plotted for SD, max, min, range, and all descriptors combined. SNR varies from -10 dB to +Inf (clean signal). Data taken from CGN.*

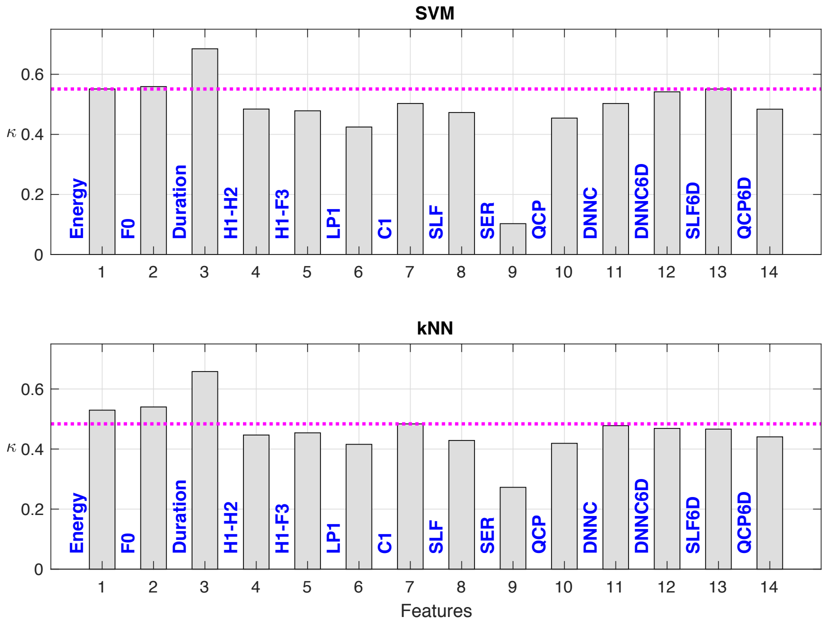

Figure 10: *SVM and kNN Fleiss kappa (κ) performance for all measures, combined descriptors, and clean speech for CGN. The horizontal dotted line marks the overall best tilt performance.*

4.2.2 C-PROM

The results for the multidimensional tilt measures for C-PROM, similarly to the case for CGN, appear to improve over their one-dimensional counterparts. An overview of the tilt measures' performance can be seen in Figures 11 and 12 for both SVM and kNN and for Fleiss kappa. Taken together, for C-PROM, the best performing tilt measures (SVM, combined descriptors) are DNNC6D (κ = 0.36), SLF6D (κ = 0.36), and QCP6D (κ = 0.34). In contrast, the corresponding one-dimensional measures (SVM, combined descriptors) attain for DNNC κ = 0.20, SLF κ = 0.15, and QCP κ = 0.14. In all, the performance of the multidimensional measures for C-PROM is close to that of energy, F0, and duration (see Figure 12).

Across the different statistical descriptors, it can be seen in Figure 11 that performance varies with the lowest scores attained for *mean* (agreement levels in the range of κ = 0 and 0.12; *mean* not plotted here separately). Conversely, the best performing descriptors for C-PROM are the *range*, *max*, and *min*. For instance, for SVM, *max*, and clean speech, DNNC6D reaches κ = 0.29, SLF6D κ = 0.24, and QCP6D κ = 0.23. The corresponding best results for the one-dimensional measures (SVM, *range*, clean speech) are for DNNC κ = 0.20, SLF κ = 0.18, and QCP κ = 0.09. For reference, the best performance for the same configuration for energy, F0, and duration is κ = 0.23, κ = 0.36, and κ = 0.38, respectively. SNR levels have a more variable impact on the performance of the multidimensional tilt measures with *max*, *range*, and *min* being affected already at approximately 15 dB SNR whereas *SD* and *mean* seem more volatile. Finally, AMR coding of speech has an impact on tilt measures with performance dropping (for SVM, *range*) to κ = 0.21, 0.15, 0.13 for AMR-NB and to κ = 0.23, 0.24, 0.14 for AMR-WB, for the measures of DNNC6D, SLF6D, and QCP6D, respectively.

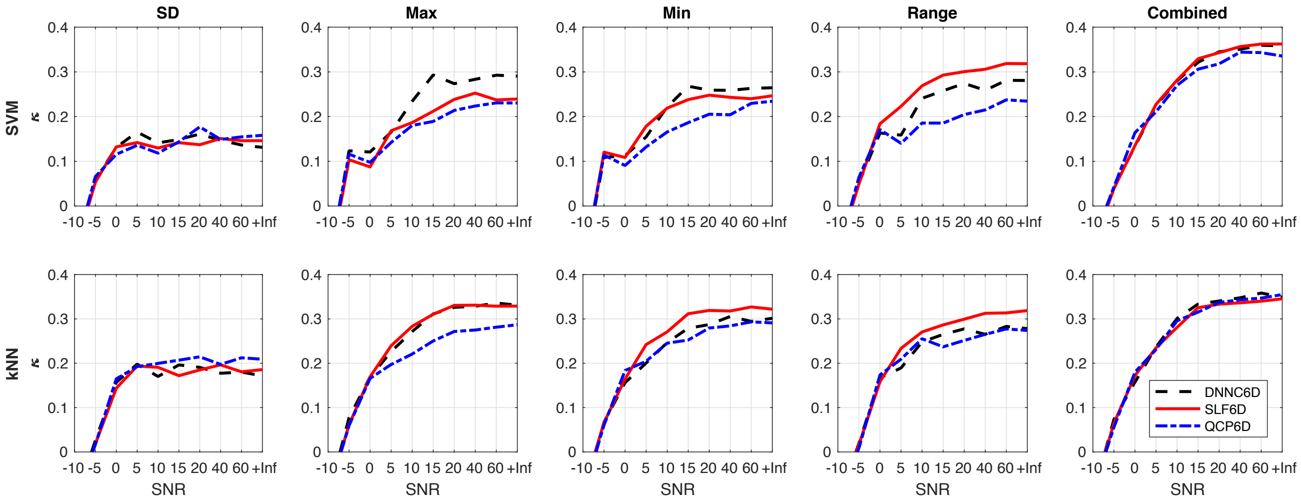

Figure 11: *SVM and kNN Fleiss kappa (κ) performance for multidimensional tilt measures plotted for mean, SD, max, min, and range. SNR varies from -10 dB to +Inf (clean signal). Data taken from C-PROM.*

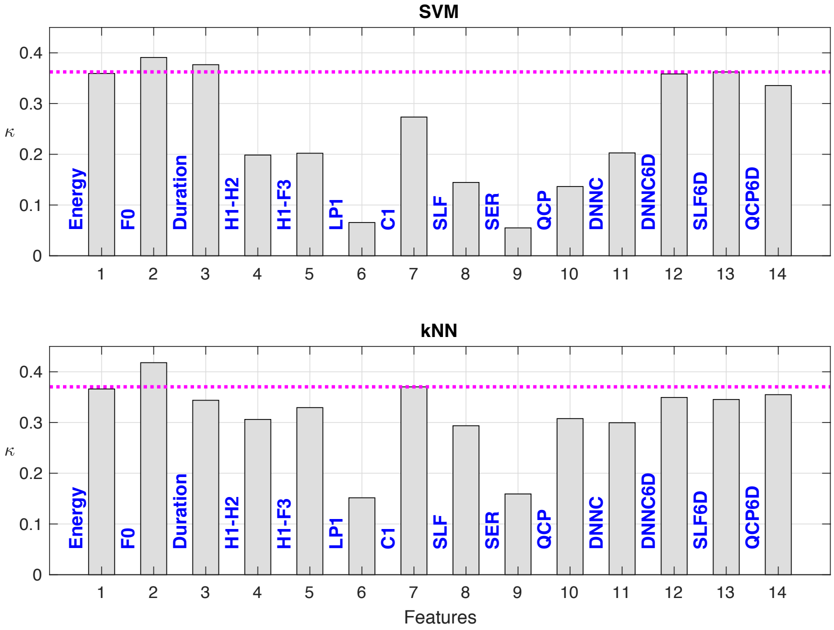

Figure 12: *SVM and kNN Fleiss kappa (κ) performance for all measures, combined descriptors, and clean speech for C-PROM. The horizontal dotted line marks the overall best tilt performance.*

4.2.3 Combinations of energy, F0, duration, and tilt

The effect of feature combinations on performance was evaluated for selected feature combinations of multidimensional and one-dimensional tilt measures together with the acoustic features of energy, F0, and duration. An overview of the results can be seen in Table 5 where all independent features and their combinations have been computed for all statistical descriptors. The best result in both corpora is achieved for the combination of energy, F0, and duration, with performance that slightly exceeds that of duration only for CGN, and with a larger difference for C-PROM. This difference in performance in the two corpora is likely due to the different role of the features in distinguishing prominence in the two languages. Specifically, in C-PROM, energy, F0, and duration may carry complementary information for prominence to such extent that almost doubles the classification result for the corpus when the features are combined. On the other hand, for CGN, the inclusion of energy and F0 in addition to duration brings a rather marginal increase in performance.

Combination of all SUT and all SOT measures does not bring improvements, and, on the contrary, reduces the performance reached when compared to the best measures of DNNC6D and SLF6D independently. This might be attributed to the observation that when combining all tilt measures that belong to the same qualitative group (SOT or SUT), even though there are observed differences in the measures, they also share much redundant information, and also information from measures that do not perform well in the task (e.g., SER or LP1). Thus, although the measures within the same group may include supplementary information for prominence, they also seem to carry much overlapping and redundant information. Similarly, when combining all tilt measures that belong to either of the two groups (SOT, SUT) together with energy, F0, and duration, performance does not improve over what is achieved with a combination of only the three (basic) features. Equally, combinations of the best performing basic features (energy, F0, duration) together with the single best performing tilt measures from the source (DNNC6D) and surface (SLF6D) groups do not improve performance beyond the result of the combined basic features. This, however, might not necessarily reflect that tilt measures do not convey relevant information for characterizing prominence. To further probe the effect of tilt in separating the two prominence categories, a combination of the best source (DNNC6D) and surface (SLF6D) tilt measures together with the features of energy and F0 was examined. The results indicated that the inclusion of tilt brings important improvements in performance in both cases and for both corpora (see also Table 5).

Table 5: *SVM classification performance for independent features and feature combinations. Data taken from CGN and C-PROM. Values in bold indicate the best performing features in each group.*

	Basic	**κ**	**Source tilt**	**κ**	**Surface tilt**	**κ**	**Combined**	**κ**	**Combined**	**κ**
CGN	F0	0.55	QCP	0.45	H1-H2	0.48	**All Basic**	**0.72**	EN+F0	0.62
	Energy	0.56	QCP6D	0.48	H1-F3	0.48	All SUT	0.52	EN+F0+DNNC6D	0.64
	Duration	**0.69**	**DNNC**	**0.50**	**C1**	**0.50**	All SOT	0.48	EN+F0+SLF6D	0.65
			DNNC6D	**0.54**	SER	0.10	All Basic+All SUT	0.68	EN+F0+C1	0.65
					LP1	0.43	All Basic+All SOT	0.66	DNNC6D+SLF6D	0.53
					SLF	0.47	**All Basic+DNN6D**	**0.71**	DNNC+C1	0.54
					SLF6D	**0.55**	**All Basic+SLF6D**	**0.72**		
	Basic	**κ**	**Source tilt**	**κ**	**Surface tilt**	**κ**	**Combined**	**κ**	**Combined**	**κ**
C-PROM	F0	0.36	QCP	0.14	H1-H2	0.20	**All Basic**	**0.60**	EN+F0	0.50
	Energy	**0.39**	**QCP6D**	**0.34**	H1-F3	0.20	All SUT	0.41	EN+F0+DNNC6D	0.53
	Duration	0.38	DNNC	0.20	**C1**	**0.27**	All SOT	0.35	EN+F0+SLF6D	0.54
			DNNC6D	**0.36**	SER	0.06	All Basic+All SUT	0.56	EN+F0+C1	0.54
					LP1	0.07	All Basic+All SOT	0.54	DNNC6D+SLF6D	0.38
					SLF	0.15	**All Basic+DNN6D**	**0.60**	DNNC+C1	0.34
					SLF6D	**0.36**	**All Basic+SLF6D**	**0.59**		

5. Discussion and conclusions

The experiments in the present work investigated the realization of prominence in speech from the perspective of the most well-known spectral tilt measures together with a recently proposed DNN-based technique, including scalar and multidimensional representations for tilt. In addition, to understand the behavior of the tilt measures under non-ideal conditions encountered in noisy real-life scenarios, corrupted versions of the original speech material were also examined. Comparisons to the widely acknowledged acoustic correlates of prominence of energy, F0, and duration were also conducted. All investigations were carried out on Dutch and French continuous speech.

The results from the present experiments revealed differences in the performance of the distinct tilt measures, indicating that the different methods also lead to different separability of the prominent and non-prominent categories. The present analysis also revealed that the inclusion of higher-dimensional parameterizations for tilt can lead to substantial performance improvements, even in the case of surface tilt where higher-order features should be already somewhat sensitive to the formant structure of speech (content vs. style). Speech degradation affected all tilt measures, where, on average, at approximately 10 dB SNR, performance started to deteriorate rapidly for all scalar measures in the added noise condition. In contrast, the multidimensional tilt measures exhibited more robustness with increasing levels of noise. Telephone-band coding of speech had a similar degrading effect on tilt, with, however, lower impact on the overall performance than low SNR additive noise. For all tilt measures examined (scalar and multidimensional) in the two corpora, both narrowband and wideband coding of speech reduced the prominence class separation with the narrowband coding always having the highest impact on performance. The most robust tilt measures across the tests for telephone-band coded speech were the DNNC6D, SLF6D, and C1.

In particular, the first experiment shows that the well-known one-dimensional measures for tilt can vary in performance, and that the well-established measures of energy, F0, and duration have an overall higher separability between prominent and non-prominent words, and appear also to be more robust in the presence of noise. In addition, it was found that the overall best performing one-dimensional measures for tilt in CGN and C-PROM were the C1, DNNC, and H1-F3, although some corpus-specific variation in performance was also evident. It is also interesting to note that the two examined corpora exhibited large differences in their class separability potential. These differences may be largely attributable to the underlying differential phonological structure of the languages and, consequently, to their inherent dissimilarities in conveying prosodic information. French and Dutch are two very distinct languages phonologically with Dutch having a clearer characterization for prominence (see, e.g., Sluijter & van Heuven, 1996a, for a discussion) and French a more intricate representation that is at times met with arguments of whether the language does carry prominence at all (see, Frost, 2011, for a discussion and comparison). This difference is also reflected in the present study where the best overall class separation for the scalar measures was substantially higher for CGN (r = 0.71 for duration) than for C-PROM (r = 0.48 for duration). This corpus-specific difference extends across all measures analyzed. However, it is difficult to disentangle language-specific differences from any other potential differences in the two corpora, such as differing recording conditions.

The second experiment aimed at investigating the potential benefits of including higher-order parameterizations of tilt in the examination of prominence class separation. The introduction of three multidimensional measures, namely, DNNC6D, SLF6D, and QCP6D, led to performance improvements in both CGN and C-PROM that exceeded that of energy for both corpora and were at approximately the same level with F0. An evaluation of all one-dimensional measures together with the multidimensional measures independently in supervised classification showed clear evidence that the multidimensional versions are among the best performing measures for tilt. Specifically, the best overall performing tilt measures were DNNC6D, SLF6D, and C1 –note that all measures in this comparison were evaluated using five statistical descriptors, therefore, the one-dimensional C1 in this case corresponds to a five-dimensional vector and, respectively, the x6D measures to 30-dimensional vectors. The performance of these measures is at the same level with energy and F0, rendering these measures as equally important for the discrimination of prominent categories in our study.

To further investigate the contribution of tilt measures and the complementary information they might add in the classification task, we also examined several feature combinations. The results showed that the best overall feature combination was that of energy, F0, and duration in both corpora whereas combinations of different tilt measures alone and together with energy, F0, and duration, did not seem to improve class separation. Although surprising, earlier studies have also observed little improvements with the addition of tilt measures in supervised classification tasks (see, e.g., Kakouros, Pelemans, Verwimp, Wambacq, & Räsänen, 2016; Streefkerk, Pols, & ten Bosch, 1999). To further examine the effect of tilt in prominence class separation performance, a combination of the best performing basic features (energy, F0, duration) together with the single best performing tilt measures from the source (DNNC6D) and surface (SLF6D) groups were tested and showed no improvements beyond what was achieved with a combination of all basic features. However, this result does not necessarily imply that tilt does not bring complementary information in characterizing prominence categories. To investigate this, combinations of energy and F0 together with DNNC6D and SLF6D were tested separately and compared to the performance of the combination of energy and F0. The result revealed improvements in class separation in both corpora indicating that tilt does bring complementary information for the task. However, the performance is still lower to that of a combination of all basic measures together. Perhaps the inclusion of tilt does not introduce much supplementary information in identifying prominence categories, especially when compared to duration, a potentially dominant feature, at least for Dutch (see Sluijter & van Heuven, 1996a).

A general observation from the findings in the present study was that some tilt measures exhibited performance that was substantially lower to that of the rest of the spectral tilt measures. In this regard, one important aspect to consider is that, although all tilt measures attempted to quantify the same superficial phenomenon, the underlying estimation procedures differed for each individual measure. For instance, it was observed that the worst performing tilt measures were the SER and LP1 in both C-PROM and CGN, where SER had the lowest overall performance. The reliance of SER on band-limited energy ratios (ratio of the spectral energies in the range between 0–1 kHz and 1–5 kHz) may make the measure more susceptible to noise present in the original signal recordings, as different noise sources have different spectral energy distributions across the frequency range. As both corpora consisted of non-ideal recordings with various noise sources being present, SER was likely greatly impacted by the noise present in the original signals. For LP1 (the first order forward linear predictor coefficient), the performance is also low for C-PROM, similarly to SER, but this is not observed for CGN where performance of LP1 is substantially better than SER. In practice, the frequency response of the first order linear prediction filter provides an approximation of tilt (similar to a regression line fit). In the case of LP1, the presence of noise can also considerably impact the estimated tilt, as, similarly to SER and any other 1-dimensional representation of tilt, LP1 can be also very sensitive to any type of noise degradation. Interestingly, measures that rely on either fitting a linear model (SLF), a near-linear approximation (LP1), or computing the energy ratio (SER) directly on the logarithmic magnitude spectrum of speech do not seem to perform well for C-PROM, whereas, in contrast, for CGN, the only measure to reach low performance level was SER. Again, this is likely related to the quality of the recordings where the overall subjective quality of CGN is better than C-PROM even though both do not comprise of high-quality speech recordings. As an example, the occasional signal clipping in C-PROM introduced unwanted harmonic frequency components, as discussed in section 4.1.1.2, and this may also impact some of the tilt estimators more than others. In comparison, the only measure that approximates the slope of the surface spectrum and performed well in both corpora, was the first MFCC coefficient, C1. Since C1 is computed as the first basis function of the discrete cosine transform (DCT), performed on the logarithmic Mel-spectrum, the used Mel-filtering may improve robustness of the measure against certain signal corruptions over the other alternatives. A further observation for C-PROM is that tilt measures that include higher order parameterizations (DNNC6D, SLF6D, QCP6D) performed systematically better. In this case, the more detailed representation of the spectrum could be more robust to noise sources that impact specific frequency regions but do not equally mask all aspects of the spectral envelope. In contrast, fitting a linear model can be very sensitive to outliers such as individual noise peaks in the spectrum.

Considering the role of surface and source tilts in prominence class characterization, measures from both categories seem to have an important impact on the separability of prominent and non-prominent words, the best variants reaching the level of basic features (energy, F0, duration). Combination of different tilt measures from the same group (source or surface) does not improve separability performance but leads to a deterioration of the overall classification result in the second experiment. In contrast, combination of the best source (DNNC6D) and surface (SLF6D) tilt leads to an increase in class separability (similarly also to other across-group combinations), indicating that the two distinct groups may hold complementary information that is relevant for prominence. This observation suggests that factors present in the glottal flow also hold a role for prominence expression, a finding that is also in line with other studies suggesting that prominence may be dependent on a number of source parameters (including F0) (see Ní Chasaide et al., 2013).

In all, the present study provides important insights about the contribution of spectral tilt to the identification of prominent words in the speech stream, including analyses under the presence of additive noise and degradations caused by telephone transmission (AMR codec). Earlier, studies have examined the impact of the narrowband and wideband versions of the AMR codec on different acoustic parameters in the speech signal (see, e.g., Guillemin &Watson, 2006; Ireland, Knuepffer, & McBride, 2015) but there has not been an in-depth examination of prosodic prominence and, also, there has been little evidence for the impact of the codec on tilt in particular. The importance of tilt in prominence has been at times elusive with its contribution being seemingly undetermined across studies, including various practices for estimating tilt either from the surface structure of the speech signal or from the estimated glottal excitation. In general, the earlier studies have not been conclusive on the role of tilt in prominence (see, e.g., Campbell & Beckman, 1997; Sluijter & van Heuven, 1996a). Our present findings suggest that both surface and source measures of tilt hold an important role in identifying prominent categories, at least for the Dutch and French speech analyzed here. However, more work is needed to further validate the present findings through the inclusion of more languages in the evaluation. In addition, in the current setup only two types of speech degradation were considered. Therefore, it would be of interest to examine more types of degradation (e.g., competing talkers) that are commonly encountered in everyday communication scenarios.

Acknowledgments

S.K. was funded by Academy of Finland projects no. 274479 and 312490. O.R. was funded by the Academy of Finland projects titled "Computational Modeling of Language Acquisition" (274479) and "Analyzing Child Language Experiences Around the World" (312105). An early version of this work was presented at the Annual Conference of the International Speech Communication Association (Interspeech-2017) at Stockholm, Sweden, in August 2017.

References

Airaksinen, M., Raitio, T., Story, B., & Alku, P. (2014). Quasi closed phase glottal inverse filtering analysis with weighted linear prediction. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, *22*(3), 596–607.

Alku, P. (2011). Glottal inverse filtering analysis of human voice production – A review of estimation and parameterization methods of the glottal excitation and their applications. Sadhana – Academy Proceedings in Engineering Sciences, vol. 36, part 5, 623–650.

Aronov, G., & Schweitzer, A. (2016). Acoustic correlates of word stress in German spontaneous speech. In *Proceedings of Tagung Phonetik und Phonologie im Deutschsprachigen Raum*.

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*, vol. *13*(2), 2–30.

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, IL.

Barbosa, P. A., Eriksson, A., & Åkesson, J. (2013). On the robustness of some acoustic parameters for signalling word stress across styles in Brazilian Portuguese. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH-2013)*, Lyon, France (pp. 282–286).

Bessette, B., Salami, R., Lefebvre, R., Jelínek, M., Rotola-Pukkila, J., Vainio, J., Mikkola, H., & Jarvinen, K. (2002). The adaptive multirate wideband speech codec (AMR-WB). *IEEE Transactions on Speech and Audio Processing*, *10*(8), 620–636.

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

Bolinger, D. (1972). Accent is predictable (if you're a mind-reader). *Language*, 633–644.

Brown, G. (1983). Prosodic structure and the given/new distinction. In A. Cutler & D. R. Ladd (Eds.), *Prosody: Models and measurements* (pp. 67–77). Berlin Heidelberg: Springer.

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

Campbell, N., & Beckman, M. E. (1997). Stress, prominence, and spectral tilt. In A. Botinis, G. Kouroupetroglou, & G. Carayiannis (Eds.), *Intonation: Theory, models, and applications (Proceedings of an ESCA Workshop)* (pp. 67–70). Athens, Greece: European Speech Communication Association (ESCA).

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).

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

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

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

Drugman, T., & Alwan, A. (2011). Joint robust voicing detection and pitch estimation based on residual harmonics. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH-2011)*, Florence, Italy (pp. 1973–1976).

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

Eriksson, A., Thunberg, G. C., & Traunmüller, H. (2001). Syllable prominence: A matter of vocal effort, phonetic distinctness and top-down processing. In *Seventh European Conference on Speech Communication and Technology (EUROSPEECH-2001)*, Aalborg, Denmark (pp. 399–402).

European Telecommunications Standards Institute (ETSI) (2011). TS 126 090: Mandatory Speech Codec speech processing functions; Adaptive Multi-Rate (AMR) speech codec; Transcoding functions. 3rd Generation Partnership Project (3GPP), Technical Release 10, Version 10.1.0 (3GPP Technical Specification 26.090, ETSI Doc. Number: TS 126 090).

European Telecommunications Standards Institute (ETSI) (2011). TS 126 204: Speech codec speech processing functions; Adaptive Multi-Rate - Wideband (AMR-WB) speech codec. 3rd Generation Partnership Project (3GPP), Technical Release 10, Version 10.0.0 (3GPP Technical Specification 26.204, ETSI Doc. Number: TS 126 204).

Fant, G., & Kruckenberg, A. (1994). Notes on stress and word accent in Swedish. In *Proceedings of the International Symposium on Prosody (ISP-1994)*, Yokohama, Japan, (pp. 124–144).

Fant, G., Kruckenberg, A., Liljencrants, J., & Hertegård, S. (2000). Acoustic phonetic studies of prominence in Swedish. *KTH TMH- Quarterly Progress and Status Report*, *41*(2-3), 1–52.

Fant, G., Liljencrants, J., & Lin, Q. G. (1985). A four-parameter model of glottal flow. *STL-QPSR*, *26*(4), 1–13.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, *76*(5), 378.

Frost, D. (2011). Stress and cues to relative prominence in English and French: A perceptual study. *Journal of the International Phonetic Association*, *41*(1), 67–84.

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

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

Gay, T. (1978). Physiological and acoustic correlates of perceived stress. *Language and Speech*, *21*(4), 347–353.

Guillemin, B. J., & Watson, C. I. (2006). Impact of the GSM AMR speech codec on formant information important to forensic speaker identification. In *Proceedings of the 11th Australian International Conference on Speech Science & Technology (SST-2006),* Auckland, Australia (pp. 483–488).

Gussenhoven, C., Repp, B. H., Rietveld, A., Rump, H. H., & Terken, J. (1997). The perceptual prominence of fundamental frequency peaks. *The Journal of the Acoustical Society of America*, *102*(5), 3009–3022.

Heldner, M. (2001). Spectral emphasis as an additional source of information in accent detection. In *Proceedings of the ISCA Tutorial and Research Workshop on Prosody in Speech Recognition and Understanding (ITRW-PSRU-2001)*, Red Bank, NJ (paper 10).

Ireland, D., Knuepffer, C., & McBride, S. J. (2015). Adaptive multi-rate compression effects on vowel analysis. *Frontiers in Bioengineering and Biotechnology*, *3*, 118.

Iseli, M., Shue, Y. L., Epstein, M. A., Keating, P., Kreiman, J., & Alwan, A. (2006). Voice source correlates of prosodic features in American English: A pilot study. In *Proceedings of the Ninth International Conference on Spoken Language Processing (ICSLP-2006)*, Philadelphia, PA (pp. 2226–2229).

Jackson, M., Ladefoged, P., Huffman, M., & Antoñanzas-Barroso, N. (1985). Measures of spectral tilt. *The Journal of the Acoustical Society of America*, *77*(S1), S86–S86.

Jokinen, E., & Alku, P. (2017). Estimating the spectral tilt of the glottal source from telephone speech using a deep neural network. *The Journal of the Acoustical Society of America*, *141*(4), EL327–EL330.

Kakouros, S., & Räsänen, O. (2014). 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. (2016). 3PRO – An unsupervised method for the automatic detection of sentence prominence in speech. *Speech Communication*, *82*, 67–84.

Kakouros, S., Räsänen, O., & Alku, P. (2017). Evaluation of Spectral Tilt Measures for Sentence Prominence Under Different Noise Conditions. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH-2017)*, Stockholm, Sweden (pp. 3211–3215).

Kakouros, S., Pelemans, J., Verwimp, L., Wambacq, P., & Räsänen, O. (2016). Analyzing the Contribution of Top-Down Lexical and Bottom-Up Acoustic Cues in the Detection of Sentence Prominence. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH-2016)*, San Francisco, CA (pp. 1074–1078).

Kane, J., & Gobl, C. (2013). Automating manual user strategies for precise voice source analysis. *Speech Communication*, *55*(3), 397–414.

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

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

Kreiman, J., Gerratt, B. R., & Antoñanzas-Barroso, N. (2007). Measures of the glottal source spectrum. *Journal of speech, language, and hearing research*, *50*(3), 595–610.

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

Lippus, P., Tuisk, T., Salveste, N., & Teras, P. (2013). Phonetic corpus of Estonian spontaneous speech. Institute of Estonian and General Linguistics, University of Tartu. DOI: https://doi.org/10.15155/TY.000D.

Lu, Y., & Cooke, M. (2009). The contribution of changes in F0 and spectral tilt to increased intelligibility of speech produced in noise. *Speech Communication*, *51*(12), 1253–1262.

Magi, C., Pohjalainen, J., Bäckström, T., & Alku, P. (2009). Stabilised weighted linear prediction. *Speech Communication*, *51*(5), 401–411.

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).

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

Murphy, P. J., McGuigan, K. G., Walsh, M., & Colreavy, M. (2008). Investigation of a glottal related harmonics-to-noise ratio and spectral tilt as indicators of glottal noise in synthesized and human voice signals. *The Journal of the Acoustical Society of America*, *123*(3), 1642–1652.

Ní Chasaide, A., Yanushevskaya, I., Kane, J., & Gobl, C. (2013). The voice prominence hypothesis: the interplay of F0 and voice source features in accentuation. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH-2013)*, Lyon, France (pp. 3527–3531).

Niebuhr, O., & Winkler, J. (2017). The relative cueing power of F0 and duration in German prominence perception. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH-2017)*, Stockholm, Sweden (pp. 611–615).

Okobi, A. O. (2006). Acoustic correlates of word stress in American English. Doctoral Dissertation, Massachusetts Institute of Technology.

Oostdijk, N. H. J., Goedertier, W., Van Eynde, F., Boves, L., Martens, J.-P., Moortgat, M., & 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, Spain (pp. 340–347).

Prieto, P., & Ortega-Llebaria, M. (2006). Stress and Accent in Catalan and Spanish: Patterns of duration, vowel quality, overall intensity, and spectral balance. In *Proceedings of Speech Prosody (SP-2006)*, Dresden, Germany (pp. 337–340).

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).

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

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

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

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

Sluijter, A. M. C., & van Heuven, V. J. (1996b). Acoustic correlates of linguistic stress and accent in Dutch and American English. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP-96)*, Philadelphia, PA (pp. 630–633).

Sluijter, A. M., van Heuven, V. J., & Pacilly, J. J. (1997). Spectral balance as a cue in the perception of linguistic stress. *The Journal of the Acoustical Society of America*, *101*(1), 503–513.

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).

Sundberg, J., & Nordenberg, M. (2006). Effects of vocal loudness variation on spectrum balance as reflected by the alpha measure of long-term-average spectra of speech. *The Journal of the Acoustical Society of America*, *120*(1), 453–457.

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

Terken, J., & Hermes, D. (2000). The perception of prosodic prominence. In M. Horne (Ed.), *Prosody: Theory and experiment. Studies presented to Gösta Bruce*, Dordrecht, The Netherlands: Kluwer, 89–127.

Terken, J., & Nooteboom, S. G. (1987). Opposite effects of accentuation and deaccentuation on verification latencies for given and new information. *Language and cognitive processes*, *2*(3-4), 145–163.

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

Turk, A. E., & Sawusch, J. R. (1996). The processing of duration and intensity cues to prominence. *The Journal of the Acoustical Society of America*, *99*(6), 3782–3790.

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.

Wagner, M., & Watson, D. G. (2010). Experimental and theoretical advances in prosody: A review. *Language and cognitive processes*, *25*(7-9), 905–945.

Yanushevskaya, I., Gobl, C., Kane, J., & Ní Chasaide, A. (2010). An exploration of voice source correlates of focus. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH-2010)*, Makuhari, Chiba, Japan (pp. 462–465).

Yanushevskaya, I., Murphy, A., Gobl, C., & Ní Chasaide, A. (2016). Perceptual Salience of Voice Source Parameters in Signaling Focal Prominence. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH-2016)*, San Francisco, CA (pp. 3161–3165).