Automatic Detection of Sentence Prominence in Speech Using Predictability of Word-level Acoustic Features

Aalto University
School of Electrical Engineering

Sofoklis Kakouros, Okko Räsänen
Department of Signal Processing and Acoustics
Aalto University, Finland

Background

Sentence prominence (or stress) can be generally defined as a property of speech where one or multiple words within a sentence receive special emphasis.

In natural conversation, for instance, it is common that speakers make some words more prominent than others in order to draw the listener's attention to those parts of the utterance that carry the most information.

In a recent study [1] it was shown that the temporal unpredictability of the *fundamental frequency (F0) trajectories* was connected with the perception of sentence prominence, thus giving support to the idea that unpredictability of the sensory stimulus is driving the listener's attention and thereby perception of prominence.

In this paper, we extend the earlier findings in [1] to a prominence detection system. We propose a method for the automatic detection of sentence prominence that does not require explicit prominence labels for training and that can capture prominent words in a manner hypothesized to be analogous to human perception.

Methods

Rationale: *mark words as prominent if the temporal evolution of the prosodic features is unpredictable during the words, violating the expectations of the listener (or the model) and thereby capturing the attention of the listener.*

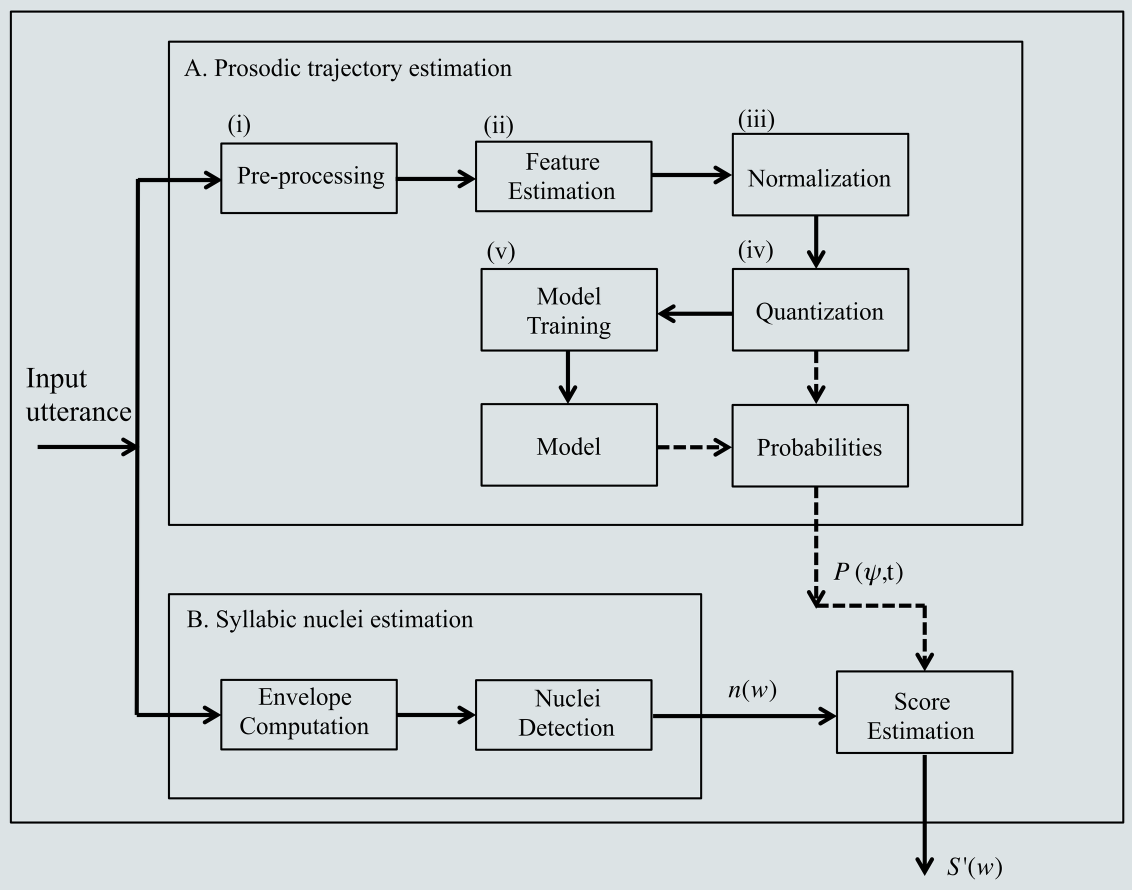

Figure 1: Overview of the processing steps during training and testing.

The proposed algorithm consists of two main blocks (see Fig. 1):
(i) a method for the detection of syllabic nuclei and
(ii) a statistical model that learns the typical prosodic trajectories.

Syllabic nuclei estimation

In order to estimate the number of syllabic nuclei in each word (or per time unit), the smoothed signal amplitude envelope is used to segment speech into subsequent syllables (see [2] for details).

Statistical modeling

Three acoustic correlates of prominence:
(i) energy,
(ii) fundamental frequency (F0),
(iii) duration (word and syllable).

For the statistical modeling of the temporal evolution of the prosodic feature trajectories, n-gram probabilities are computed from the relative frequencies of different n-tuples of quantized features in the training data.

In order to measure the overall predictability of the prosody during each word, F0 and energy-based word-level prominence scores are computed for each word by integrating the instantaneous feature probabilities over the duration of the entire word.

Each acoustic feature based word score in the utterance is weighted by the exponent of the average syllable duration (see Fig. 2) for the non-linear mapping of the durational nucleic information.

The prominence classification for each word is then determined based on whether the word-level score falls below a threshold.

Experiments and Results

Speech Material: CAREGIVER Y2 UK corpus [3]:
- testing: 300 prosodically annotated utterances (≈ 30 minutes of data) from one male and female talker (600 in total),
- training: 9594 utterances from 9 speakers ≈ 7.2 hours of data.

Data Collection: twenty subjects (11 male, 9 female, age range 20-61 with a median of 30 years) marked the perceived prominence in the test set.

Evaluation: two evaluation approaches computed between algorithmic output and human annotations:
(i) the standard Fleiss Kappa statistic,
(ii) Precision, Recall, F-Score, and Accuracy.

Experiment: the experiment was run in a cross-validation setup where data from 9 speakers were used for training and one for testing. Three orders of the n-gram model (n = 2, 3, and 4) were used for training and testing was carried out on the held-out set of 300 annotated utterances (on one of the two annotated speakers). None of the test signals were used in training.

Results: overall, the algorithmic output converges with the annotators' prominence responses with 86% accuracy:
- in terms of the individual features' performance, F0 (ACC=86.20%) and energy (ACC=86.16%) seem to be equally descriptive in determining prominence,
- syllable duration alone has much lower F and kappa measures, indicating that independently it does not explain prominence as accurately (see also Table 1),
- several feature combinations were also tested and the best performance was achieved for energy and F0 (ACC=86.95%).

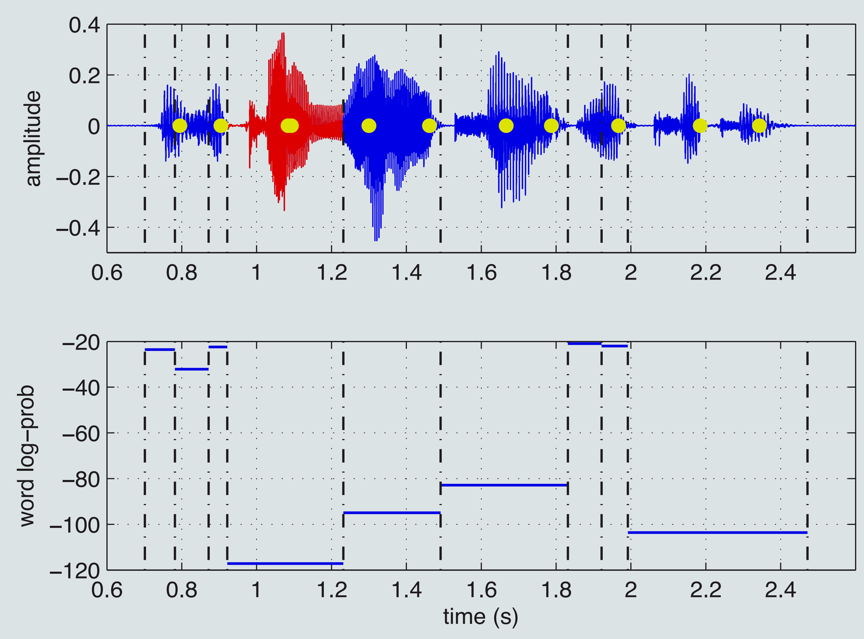

Figure 2: Example output of the algorithm for the utterance "There is a clean yellow cow and a cookie". Top panel: original signal waveform where red marks the word perceived as stressed by the majority of the listeners and yellow marks the syllabic nuclei. Bottom panel: word scores produced by the proposed algorithm.

Table 1: Prominent word detection performance for the individual features and their combination (for λ = 0.7) pooled over the three n-gram orders (n = 2, 3 and 4) and averaged across speakers.

	ACC	F	PRC	RCL	Fleiss Kappa
F0+EN	86.95% ±0.18	71.99% ±0.39	73.25% ±0.13	70.78% ±0.87	0.61 ±0.04
F0	86.20% ±0.39	71.22% ±0.99	73.60% ±1.01	69.00% ±0.97	0.60 ±0.01
EN	86.16% ±0.34	70.15% ±0.57	71.02% ±1.21	69.31% ±0.05	0.59 ±0.09
Syllable duration	80.72% ±0.15	53.90% ±0.10	56.85% ±0.17	51.23% ±0.10	0.37 ±0.01

Conclusions

A new and effective approach to automatic prominence detection based on the predictability of prosodic trajectories.

The results for the best feature combination show accuracy of 86.95% with the annotators' responses providing initial promising results for the method.

This level of performance compares well with other approaches that do not use prosodic labels.

References

[1] S. Kakouros and O. Räsänen, "Statistical unpredictability of F0 trajectories as a cue to sentence stress," in Proceedings of the 36th Annual Conference of the Cognitive Science Society, pp. 1246–1251, Quebec, Canada, 2014.
[2] O. Räsänen, G. Doyle, and M. C. Frank, "Unsupervised word discovery from speech using automatic segmentation into syllable-like units," in Proceedings of Interspeech, Dresden, Germany, 2015.
[3] T. Altosaar, L. ten Bosch, G. Aimetti, C. Koniaris, K. Demuynck, and H. van den Heuvel, "A speech corpus for modeling language acquisition: CAREGIVER", in Proceedings of the International Conference on Language Resources and Evaluation (LREC), pp. 1062–1068, 2010.

Acknowledgements

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