

AN UNSUPERVISED QUANTITATIVE MEASURE FOR WORD PROMINENCE IN SPONTANEOUS SPEECH

Dagen Wang, Shrikanth Narayanan

Speech Analysis and Interpretation Lab
USC Viterbi School of Engineering, Los Angeles, CA 90089

ABSTRACT

An unsupervised approach for automatic speech prominence detection is proposed in this paper. The algorithm scores prominence by fusing different acoustic feature sets from the speech signal correlation envelope. In addition, we investigate part of speech (POS) as a linguistic correlate for speech prominence. We also underscore the inadequacy of the traditional approach to prominence detection of heuristically tagging speech prominence into discrete levels (categories). Instead, we propose to keep the prominence score continuous, evaluate it by correlation with POS, and leave it for further processing by other applications such as natural language understanding. Furthermore, in contrast to most previous studies, we evaluate prominence scoring on spontaneous speech data (switchboard corpus). Our experimental results indicate that the proposed prominence score can robustly distinguish between content word and function word classes.

1. INTRODUCTION

Speech prominence detection is useful in many speech applications. In most natural language understanding (NLU) systems knowing only the speech-to-text transcription by itself are not sufficient. Prominence, which refers to a prosodic property, and not directly conveyed by transcription, is valuable in "understanding" the speech. For instance, previous work [20] has demonstrated its usefulness in clarifying the ambiguities in specific utterances. This is gaining increasing importance in recent years as more and more effort is directed toward spontaneous speech processing. The greater acoustic variability found in this realm makes acoustic modeling more challenging. Similarly, disfluency and the incomplete syntax prevalent in spontaneous speech degrade the traditional language modeling. Hence as a result, automatic recognition of spontaneous speech is considerably worse translating to serious problems for most of NLU algorithms. However, it is apparent that prosodic events carry rich and critical information in spontaneous speech communication. Hence, information about prominence, as well as other prosodic events, can play a more important role in processing and understanding spontaneous speech. For example, locating content words is an important goal in NLU and its accuracy has a crucial influence on the overall system performance. So, in addition to the semantic analysis of the text, measures of speech prominence could serve as a useful feature in this task of content word location.

1.1. The notion of prominence

The notion of the prominence seems not be well defined. Terken defines it as words or syllables that are perceived as standing out from their environment [5], or in other words, it refers to the perceptual salience of a language unit [3]. It also does not have clear boundary against sentence accent nor pitch accent [2]. We will soon see that many research challenges arise from the fuzzy nature of this definition.

1.2. Prominence measure

In order to enable an objective study on this problem, prominence has to be quantified. Various methods have been proposed each with its own advantages and disadvantages. Almost all researchers tend to take the approach of imposing discrete categorization on prominence measures. For instance, Portele and Heuft [7] define prominence on a scale from 0 to 30 at the word level. But this measure is a challenging, if not impossible, task for the human transcribers that provide data for model training. On the other hand, a relatively easy task is just marking if a word is prominent or not [3] i.e., the prominence level of each word can be marked as either 0 or 1. It should be noted that in these, and most other previous studies, people have mostly used read speech as the data source. Even for those data, studies show that people only reach limited agreement (81% in [4], 87% in [8]) on word level prominence annotation.

In this paper, we propose a new method for scoring prominence on a continuum that especially targets spontaneous speech. It combines spectral and temporal features. Further, correlation of this prominence score to a linguistic measure (part of speech) is investigated rather than attempting classification into discrete prominence levels. Hence, the algorithm does not rely on manual transcription of prominence levels but does utilize speech-to-text information from ASR/transcriptions. The rest of the paper is organized as follows: Sec 1.3 discusses the acoustic correlates of prominence; Sec 1.4 introduces part of speech as a prominence correlate. Sec 2 describes the algorithm to get the syllable nuclei, generate the prominence score and POS-based evaluation. Sec 3 discuss the experimental results. Further discussions and conclusions are provided in Sec 4.

1.3. Acoustic correlates of prominence

Numerous studies have been carried out on acoustic analysis of prominence in the recent several years and there is a rough agreement in the literature that syllable duration, pitch tilt, and intensity (or sub-band energy) have close correlation with

speech prominence. Nevertheless a number of key questions remain unanswered including: Are all these proposed features equally important and if not, what is their relative importance? Are there other possible features that better correlate with prominence? Finally, the difficulties in transcribing prominence easily and reliably make it difficult in directly applying well established supervised machine-learning methods.

1.3.1. Syllable duration

Syllable duration is a straightforward feature. Speakers tend to stretch the constituent syllable durations when they try to emphasize a specific word [13]. Recent research trends in locating syllable boundaries have moved from knowledge-based to statistics based approaches [9]. But neither of these methods seem to provide satisfactory performance on continuous read speech, let alone spontaneous speech. Due to these difficulties in precisely locating syllable boundaries, we have instead opted for using speech rate information. Specifically, we adapt and modify the algorithm proposed for speech rate estimation in [17]. Details are provided in Section 2.

1.3.2. Pitch feature

Pitch patterns have been shown to have strong correlation with prominence. Again, several efforts have been made to quantitatively describe the complex pitch trajectory behavior. One trend is trying to enumerate all possible patterns by using a multi-level profile description [10]. A famous example of this is the TOBI system [11]. Yet this approach has its problem in its self-completeness (and as a result many prosodic systems are using a variant system modified from TOBI to meet specific needs [22]) and transcribing effectiveness. The other trend is to sacrifice or distort details of the pitch behavior and extract the very key features. A widely used approach is the Rise/Fall/Connection (RFC) model [12]. Nevertheless, almost all of these approaches treat pitch as a suprasegmental feature. What we do in this paper is to consider the pitch behavior in the syllable nuclei range. As the pitch range in this segmental range decreases considerably, we hypothesize that a simple modeling scheme would work in capturing the pitch behavior. Our experiment supports this argument. (See Sec 2.2 and Sec 3)

1.3.3. Spectral intensity

Spectral intensity also correlates with prominence [1]. Research shows that band energy in 500Hz – 2000Hz has maximum correlation with prominence [13]. Note that this also approximately coincides with the sonorant band (300-2300 Hz) in Strom’s study [14]. Interestingly, the other 2 bands ([0-300] Hz, [2300-6000Hz]), related to the nasal and fricative bands, have been shown to have not much acoustic correlation with prominence. In this paper, we focus on the sonorant band but apply with both a temporal and spectral correlation method.

1.3.4. Fusion methodology

In this paper, we utilize the aforementioned features, but both the feature processing and the fusion approach are different from other previously published work. The motivation of this algorithm comes from two sources: Empirical; Learning from development test.

1.4. Part of speech measure & performance evaluation

For evaluation, people have used different quantization (categorization) methods to transcribe prominence (Section 1.3). The inherent uncertainty in any manual classification algorithm leads to the danger to lose and distort the original information. This point is underscored within the fuzzy logical community [15]. Yet in order to evaluate, we need a measure of prominence within a common, consistent reference. In this paper, we propose to use POS measure to provide the reference framework.

POS has been long and very well studied within the NLU community. It might be viewed as a shallow parsing of language. Even though it is far from enough to convey the meaning of the language, it carries adequate information to convey speech prominence. For instance, people tend to be prominent on content words compared to functional words. Even though there may be exceptions to this rule, those cases are deemed to be statistical insignificant

Another advantage for using POS measure is that automatic POS tagging has very good performance. Baseline unigram systems have the accuracy of about 90%. In Brill’s tagger, that uses simple contextual knowledge, the performance has reached 97.2% [16]. However, we should note that the tagging performance using automatically transcribed spontaneous speech may be somewhat lower but still adequate for our purposes. We use Brill’s tagger in this work.

2. ALGORITHM DESCRIPTION

Based on the aforementioned discussion and ideas, this section tries to systematically describe the proposed algorithm in 3 parts. Part1 describes the approach for obtaining the spectral-temporal correlation envelope and the syllable nuclei. Part2 describes the fusion algorithm for the prominence score. Part3 describes how we evaluate the results.

2.1. Syllable nuclei location

Figure 1 is the illustration of the key component of this algorithm. The algorithm extracts the syllable nuclei from the correlation envelope of the speech signal. We extend the idea of sub-band correlation [17] and combine it with temporal correlation for getting the correlation envelope. In [17], a point-wise correlation between pairs of compressed sub-band energy is computed to get the envelopes. Further peak counting is performed to get the underlying syllable numbers. In our approach, the following modifications are made:

1. As described in Section 1.3.3, the sonorant band is the most informative in the context of prominence detection. So instead of performing full band analysis we only concentrate on the sonorant band.
2. We make finer, and more, band divisions. They are Butterworth band-pass filters centered at 360 | 480 | 600 | 720 | 840 | 1000 | 1150 | 1300 | 1450 | 1600 | 1800 | 2000 | 2200 [20]. The purpose is to track formant movement by selecting high energy subbands to do correlation (refer [6] for details).
3. Instead of doing a point-wise correlation, we do a selected sub-band correlation. Using experimental results from a development test set, we choose the 3 bands that have most energy and correlate them. This is consistent with the formant properties of vowel sound.
4. In order to make syllable nucleus more apparent and make the final envelope smooth, not only is correlation performed

spectrally, but also temporally. Here again, as a result of experiments on the development test, a temporal 70ms window (7 frames) was chosen for correlation (further details are in [6]).

5. To counter spurious peaks in the correlation envelope due to fricatives and other non-speech noise, pitch verification is introduced to the algorithm. Peaks corresponding to unvoiced portions are rejected. (refer to figure 1(b))

6. Even with all the aforementioned processing steps, there may be still some noise in the correlation envelope that makes simple peak counting not a robust measure of syllable count. To counter this issue, further smoothing techniques are applied to the envelopes. One is by applying a Gaussian smoothing window, and the other is by setting a minimum threshold on peak heights. Again, these parameters are learnt using a development test.

After these refinements to the algorithm, we get a final correlation envelope of the speech signal. In addition to peak counting, we also keep the information of the bottoms (local minimum of the correlation envelope). Each bottom-peak-bottom could represent a syllable nucleus. And the bottom-bottom distance provides an estimate of the syllable duration.

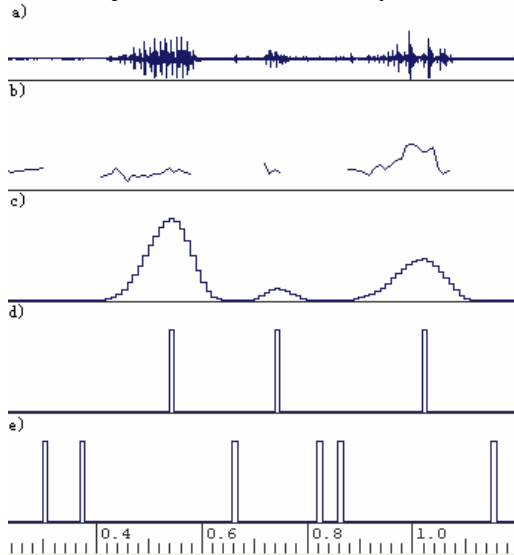


Fig1. Illustration of algorithm described in Section 2.1
a) speech waveform. b) pitch. c) correlation envelope
d) maximum(peaks) e) minimum(bottoms)

2.2. Prominence score computation

Now with the correlation envelope, bottom-peak information and pitch information, we generate the prominence related features and finally compute the prominence score. As mentioned in Sec 1.2, we assume quantization of prominence levels. For each syllable nuclei (bottom-peak-bottom):

1. The duration score is retrieved by computing the bottom-bottom distance and normalizing it by the maximum duration. Since we apply the pitch verification and minimum thresholding for peaks, there exist, albeit infrequently, cases that have no measurable peak between neighboring bottoms. So the distance we consider here is the neighboring bottoms that have a peak in between. (refer to figure 1(d) and 1(e))

2. The spectrum score is represented by the peak value normalized by the maximum peak. The peak comes from the selected sub-band correlation and the temporal correlation (see

Sec 2.1). This is our choice for spectrum intensity representation. We do not consider the integration of the envelope [1], since the duration factor will be recomputed.

3. By median filtering and jump removal of pitch curve, we use median pitch inside a syllable nucleus to score the pitch feature. Of course, normalization by maximum pitch is applied.

We choose an unsupervised approach. The final score is the mean of these 3 scores. (refer Sec 4 for further discussions.)

2.3. POS evaluation

Now we need to map the syllable score to the word, then to the POS. The word transcriptions can come from an ASR decoder or the manual transcription from the data provider (see Sec 3).

1. We use Brill's tagger in this task. The word transcription is passed to the tagger and the POS taggings are recorded.

2. Tag the word boundary timing information (from the data transcription, see Sec. 3) on the correlation envelope.

3. Distribute the syllable nuclei to its belonging word according to the timing information in step 2.

4. Such a word might have multiple syllables. Based on observations by other researchers, and our own, we can note that speakers tend to express the prominence of a word in the stressed syllable. So we keep the syllable with the highest prominence score to represent the word's prominence and discard all other syllables.

5. From step 1, each word has its POS tagging. Now we treat the word score in step 4 as its POS score. Then record this POS score in the POS hash tables. For example, "book" has a score 0.6, "table" has a score 0.7, these two word contribute the "NOUN" in the POS hash table with two items of 0.6 and 0.7.

6. Now we can get the final POS score by averaging all the items in each POS category. Then, we can compare the POS scores of functional words and content words. The difference will indicate the performance of the prominence detection algorithm.

3. DATA AND RESULT

The data used are from 5757 utterances found in the Switchboard corpus, comprising approximately four hours of data. These utterances were phonetically hand transcribed by linguists in the Switchboard Transcription Project at ICSI [18]. In order to make the algorithm general, we compute the syllable nuclei by our own method. However, for these experiments we used the word level transcriptions made available as input to the POS tagger. (See Sec 2.3)

The following components are computed by the speech filing system: sub-band energy computation, pitch estimation, median filtering and jump removing [19].

We randomly selected 315 spurts as a development set. Temporal correlation window length, sub-band number, Gaussian smoothing window, minimum thresholding are tuned with respect to the differences between the measured syllable number and transcribed syllable number. With all these processing, the final correlation envelope is retrieved.

Then, prominence score computation and POS evaluation were implemented as described in Sec 2.2 and Sec 2.3.

Figure 2 shows a plot of prominence scores for different POS categories. The left 4 represent functional word category and the right 4, content word category. Ideally, we desire a statistically significant difference between these 2 classes. ANOVA analysis

performed on these 2 classes indicates that was indeed the case. We get a p value < 0.0001.

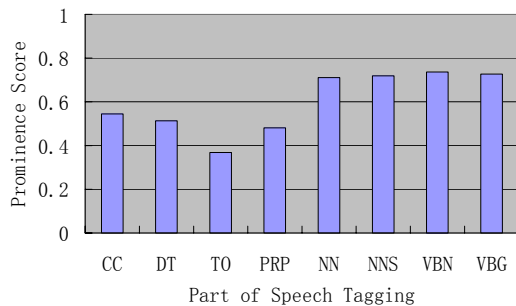


Fig2 Result POS-based prominence scores: CC:conjunction, DT:determiner, TO: "to", PRP:preposition, NN(single noun), NNS(plural noun), VBN:verb(past), VBG:verb(gerund).

4. CONCLUSIONS AND DISCUSSION

It is apparent from the results (Fig.2) that content words have higher prominence scores than functional words (by about 20% in the maximum scale of 1). There is a "step" between the 2 classes. These results were computed from 5757 spontaneous speech utterances, and attest to statistical generalizability of the observation. In this sense, we can note that the proposed prominence score conveys useful information in processing spontaneous speech.

In Section 2.2, we obtained the prominence score as a mean of three different measures. In general, this may not be optimal. Optimal score combination through (supervised) machine learning may provide a more principled approach. There is of course a trade-off between supervised and unsupervised methods. Supervised approach for learning needs more data transcription and reference prominence score (through some discretization) that may be unreliable. Further more, the learning algorithm (e.g., CART, ANN and SVM) is far from mature to optimally "understand" the feature patterns. That is one reason why the unsupervised approach is favored in this work since it is fast and straightforward.

The other possible advantage of this evaluation is the data selectivity. We mentioned that manually transcribed prominence has limited word level agreement [4][8]. Our approach does not require direct prominence transcription but attempts to infer its utility indirectly from linguistic information such as content word detection.

There are other applications beyond that illustrated. For instance, the prominence score could also work as a confidence score for automatic speech recognition. Such scores can be used in conjunction with language models to appropriately weight lexical items e.g., function versus content words since they tend to have different discrimination behavior (function words are more error prone at decoding). Similarly prominence scores can be used in automatic NLU to improve its performance. An example application is what we are doing in MRE project [21]. Details of such applications will be described in future work.

5. REFERENCES

- [1] Tamburini, F. "Automatic Prosodic Prominence Detection in Speech using Acoustic Features: an Unsupervised System", *Proc. Eurospeech 2003*, Geneva, 129-132.
- [2] Streefkerk, B.M., "Acoustical correlates of prominence: a design for research", *Proc. Inst. of Phon. Sciences*, Vol.20, 1997.
- [3] Streefkerk, B. M. et al., "Acoustical features as predictors for prominence in read aloud Dutch sentences used in ANN's. ", *In Proc. Eurospeech '99*, Budapest, pp. 551-554, 1999.
- [4] Pitrelli, J.; Beckman, M.; Hirschberg, J., 1994. "Evaluation of Prosodic Transcription Labeling Reliability in the ToBI framework. " *In Proc. ICSLP*, vol 2, 123-126
- [5] Terken, J., 1991. "Fundamental frequency and perceived prominence of accented syllables." *J. Acoust. Soc. Am.*, 89.
- [6] D. Wang, S. Narayanan, "Speech Rate Estimation via temporal correlation and selected subband correlation", submitted to *ICASSP 2005*.
- [7] Portele, T. and Heuft, B., 1997. "Towards a prominence-based synthesis system". *Speech Communication*, 21, 61-71,
- [8] Grice, M.; Reyelt, M.; Benzmuller, R.; Mayer, J.; Batliner, A., 1996. "consistency in Transcription and Labeling of German Intonation with GToBI. " *In Proc. ICSLP*, vol. 3, 1716-1719.
- [9] Howitt, A.W., "Automatic Syllable Detection for Vowel Landmarks", *PhD Thesis*, MIT, 2000.
- [10] Campione, E. and Veronis, J., "A multilingual prosodic database", *In Proc. ICSLP98*, Sydney, 1998.
- [11] J. Pierrehumbert, and J. Hirschberg. "TOBI: A standard for Labeling English Prosody". *In Proceedings ICSLP 1992*.
- [12] Taylor, P.A., "The rise/fall/connection model of intonation." *Speech Comm.*, 15, pp. 169-186, 1995.
- [13] Sluijter, A. and van Heuven, V., "Acoustic correlates of linguistic stress and accent in Dutch and American English. ", *In Proc. ICSLP96*, Philadelphia, pp. 630-633, 1996.
- [14] Strom, V. "Detection of accents, phrase boundaries, and sentence modality in German with prosodic features," *Proceedings Eurospeech95*, Madrid, Vol. 3: 2039-2041.
- [15] Mendel, Jerry M., 1995: "Fuzzy Logic Systems for Engineering: A Tutorial", *Proceedings of the IEEE*, vol. 83, no. 3, March 1995, pp. 345-377
- [16] Eric Brill, "A report of recent progress in transformation-based error-driven learning", *AAAI 1994*
- [17] N. Morgan and E. Fosler-Lussier. "Combining multiple estimators of speaking rate". *In IEEE ICASSP-98*, Seattle, WA, May 1998
- [18] Greenberg, S., "The Switchboard Transcription Project, " in F. Jelinek, editor, *1996 Large Vocabulary Continuous Speech Recognition Summer Research Workshop Technical Reports*, chapter 6. Center for Language and Speech Processing, Johns Hopkins University, April 1997. Research Notes No. 24.
- [19] Speech filing system, <http://www.phon.ucl.ac.uk/resource/sfs>
- [20] Beckman, M.E. and Venditti, J.J., "Tagging prosody and discourse structure in elicited spontaneous speech." *In Proc. Science and Technology Agency Priority Program Symp. on Spontaneous Speech*, Tokyo, pp. 87-98, 2000.
- [21] Hill, R.W. et al. "Virtual Humans in the Mission Rehearsal Exercise System". *KI on Embodied Conversational Agents 2003*.
- [22] Wightman, C.W., "ToBI Or Not ToBI?", *speech prosody 2002*, France, 11-13 April 2002