An Acoustic Measure for Word Prominence in Spontaneous Speech

Dagen Wang and Shrikanth Narayanan, *Senior Member, IEEE*

***Abstract*—An algorithm for automatic speech prominence detection is reported in this paper. We describe a comparative analysis on various acoustic features for word prominence detection and report results using a spoken dialog corpus with manually assigned prominence labels. The focus is on features such as spectral intensity and speech rate that are directly extracted from speech based on a correlation-based approach without requiring explicit linguistic or phonetic knowledge. Additionally, various pitch-based measures are studied with respect to their discriminating ability for prominence detection. A parametric scheme for modeling pitch plateau is proposed and this feature alone is found to outperform the traditional local pitch statistics. Two sets of experiments are used to explore the usefulness of the acoustic score generated using these features. The first set focuses on a more traditional way of word prominence detection based on a manually-tagged corpus. A 76.8% classification accuracy was achieved on a corpus of role-playing spoken dialogs. Due to difficulties in manually tagging speech prominence into discrete levels (categories), the second set of experiments focuses on evaluating the score indirectly. Specifically, through experiments on the Switchboard corpus, it is shown that the proposed acoustic score can discriminate between content word and function words in a statistically significant way. The relation between speech prominence and content/function words is also explored. Since prominent words tend to be predominantly content words, and since content words can be automatically marked from text-derived part of speech (POS) information, it is shown that the proposed acoustic score can be indirectly cross-validated through POS information.**

***Index Terms*—Part of speech, prominence detection, rich speech transcription, spoken language processing.**

I. Introduction

SPEECH prominence detection is useful in many spoken language applications. To create a more complete natural language understanding (NLU) capability, knowledge of just "what" was spoken, as provided by speech-to-text transcription, by itself is not sufficient; knowledge about "how" a message is communicated i.e., information about linguistic and affective expressions is also important. Prominence, which refers to a prosodic property, and information that is typically not directly conveyed by conventional speech recognition systems, provides valuable linguistic information key to understanding speech. For instance, previous work [21] has demonstrated the usefulness of measures of prominence in clarifying ambiguities in spoken utterances. Another motivation for automatic prominence detection arises in the context of robust processing of spontaneous speech, a task rendered difficult due to several factors. The greater acoustic variability found in spontaneous speech makes acoustic modeling more challenging. Similarly, the incomplete syntax and disfluency prevalent in spontaneous speech degrade the power of traditional N-gram language models. Hence, as a result, automatic recognition of spontaneous speech is considerably more difficult, in turn leading to serious problems for things that follow such as NLU algorithms that operate on the speech recognizer's output. However, it is well known that prosodic events, often ignored in ASR, carry rich and critical information in spontaneous speech communication. Hence, one could expect that information about prominence, as well as other prosodic events, can play an 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 derived from the acoustic signal could serve as a useful feature for automatic content word location.

A. The Notion of Prominence

There is a diversity of viewpoints in defining the notion of prominence. Terken defines prominence as "words or syllables that are perceived as standing out from their environment" [5], while Streefkerk *et al.* use it to refer to the "perceptual salience of a language unit" [3]. It also does not have a clear distinction with respect to sentence accent or pitch accent [2].

Although there exist subtle differences and many debates regarding these terms, we will not try to survey them in this paper. Instead, we focus on the challenges of discretely categorizing speech data directly into such perceptual terms. We point out that many challenges arise from the fuzzy nature of the description of prominence, and offer one solution to indirectly score word-level prominence.

Prosody can be viewed at two levels: one, which is at the physical signal level, is explored through acoustic phonetic properties such as pitch, duration, and intensity. The other, at a more abstract level, focuses on its role within the abstract linguistic structure rather than its physical realization. Prominence, considered as a prosodic attribute, thus could habitually be viewed within these two scopes [24]. Moreover, most definitions of prominence are inherently perceptually motivated with

Manuscript received November 9, 2005; revised March 23, 2006. This work was supported by a National Science Foundation CAREER award, an ONR/MURI award, and grants from the National Institute of Health, Department of the Army, and the Defense Advanced Research Projects Agency. A preliminary version of this work was presented at ICASSP 2005 and was invited for submission to the Transactions. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Gerland Schuller.

The authors are with the Viterbi School of Engineering, University of Southern California (USC), Los Angeles, CA 90007 USA (e-mail: wangdagen@yahoo.com; shri@sipi.usc.edu).

Digital Object Identifier 10.1109/TASL.2006.881703

its physical realization linked to the speech production process. Unfortunately, there is much unknown about the underlying perception and production details, and few quantitative studies have been done [27].

The subjective nature of prominence, coupled with the lack of clear categorization, poses challenges for deriving a reliable quantitative measure to describe it. Specifically, this underscores two challenges for devising any scheme for automatic prominence description: First, depending on the chosen representation of prominence, human perception can be greatly variable among people and hence in turn impacts the use of such information for automated learning and or evaluation. Second, the relations between prominence and acoustic-phonetic speech characteristics and other higher level linguistic structure are not completely understood, as evidenced from a number of prior studies. In this paper, we consider both challenges: investigation of various acoustic measures that better correlate with prominence, as well as finding efficient ways of automatically detecting and validating prominence.

B. *Quantifying Prominence*

In order to enable an objective study on this problem, prominence has to be quantified. Various methods have been proposed in the past, each with its own advantages and limitations. 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. Likewise, Terken [24] defines prominence on a scale from 0 to 10. But implementing this measure is a challenging task for the human transcribers that provide data for model training. On the other hand, a relatively easy task is to just mark if a word is prominent or not [3], i.e., the prominence level of each word can be tagged on binary scale of 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 humans only reach limited agreement (about 80% [4]) on word-level prominence annotation. These often serve as either training samples or for evaluating results of automatic prominence detection.

In this paper, we consider a new method for scoring prominence on a continuum. The score combines spectral and temporal speech segmental features along with prosodic features. In addition to using manually tagged data, we investigate a new indirect way for evaluating this prominence score. Specifically, the correlation of this prominence score to a linguistic measure (part of speech) is investigated as an alternative to directly attempting classification into discrete prominence levels. In this case, the algorithm does not rely on manual transcription of prominence levels but utilizes more easily available speech-to-text information from ASR/human transcriptions for modeling and validation.

The rest of the paper is organized as follows: Section I-C reviews previously reported prominence detection methodologies. Section I-D introduces part of speech (POS) as a prominence correlate. Section I-E discusses various acoustic correlates of prominence. Section II introduces the data we use in this study. Section III describes the algorithm to detect syllable nuclei, a key component of the prominence measurement, and one that does not require automatic speech recognition. Section IV provides a discriminative study on different feature categories. Section V evaluates both the supervised and unsupervised prominence score calculation on a manually tagged prominence corpus. Section VI discusses POS-based validation of the prominence scores through discrimination of content and function words. Finally, we provide our conclusions and some discussions in Section VII.

C. *Prominence Detection Methodology*

There is a long history of research studying acoustic correlates of prominence and sentence accents [26]. Focused phonetic studies often restrict the problem to tens of utterances and apply various careful labeling schemes to derive inductions on prominence cues based on empirical evidence [5], [28]. Moreover, in order to control variability, these experiments are restricted to simple utterances, sometimes involving only one or two pitch accents [24]. These studies have provided valuable insights, serving as a basis for much of the work in this domain, although more detailed investigations are required to further validate the generality of these results and to expand their scope. For example, it remains unclear how such findings will generalize when more complicated contexts are introduced and when spontaneous, as against read, speech is involved.

A majority of the engineering studies, on the other hand, tend to use automated statistical approaches on significantly larger amounts of data to support various findings [1], [3]. A recent, and a good representative, study is reported in [34], where a subset of the TIMIT utterances composed of 7327 syllables taken from 485 utterances spoken (read) by 51 different speakers of American English are used. These syllables are manually transcribed as prominent or nonprominent. The correct rate is 80.6% with 7.22% false alarm rate and 12.17% missed deletion rate. An earlier work [35] used 453 utterances from the English language ATR conference-registration dialogues as the database. These data are different from the previous ones in the sense that they are from a conversational domain (though not spontaneous). Again, those data were manually transcribed with syllable stress and accent. A 61.6% correct rate was achieved on three-class discrimination (accented/stressed but unaccented/unstressed).

In most studies, speech with manually tagged prominence is used for evaluation. This approach is considered in this paper as well. However, since manual transcription is not only an expensive and time consuming process, but it also has large unavoidable variability, and data made available in this way tends to be limited [1], [7], [24]. We hence also consider an alternative way to evaluate acoustic measures of prominence, which is described in the next section, by correlating with function/content word class discrimination. The latter is derivable through automatic POS tagging and can potentially enable working with larger amounts of data.

D. *POS and Prominence Performance Evaluation*

POS has been well studied within the natural language processing community. It might be viewed as a shallow parsing of language. Even though it is far from being adequate in conveying the meaning of the language, we argue that it carries salient information conveying speech prominence. For instance, people tend to be more prominent on content words compared to function words. This property in fact is exploited in many text-to-speech systems that rely upon the simple function/content word distinction [27]. In such systems, function or closed class words, such as prepositions and articles, are less prominent (or deaccented), while content, or open class words, such as nouns and verbs, are prominent (or accented) [27].

One of goals in this paper is to exploit POS information for evaluating prominence detection. The advantage is that it would avoid the human subjectivity in direct prominence transcription and hence is easier to scale up to using large amounts of spoken language data. Another key advantage for using POS measure is that the state of the art automatic POS tagging has very good performance. Baseline unigram systems have the accuracy of about 90%. With 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 might be somewhat lower but still adequate for our purposes. We use Brill's tagger in this work.

We should also note that there are some limitations of POS-based prominence analysis. First, there are variations of prominence realization within each word class (their behavior might be unknown). Second, the relation between expected prominence with POS has exceptions [29]. For example, some function words might have higher prominence to address special meanings.

In this paper, we consider both manually tagged prominence labels as well as the POS information for investigating automatic prominence detection.

E. Acoustic Correlates of Prominence

Numerous studies have been carried out on acoustic analysis of prominence in the recent years and there is a rough agreement in the literature that syllable duration, pitch pattern, and intensity (or subband energy) have close correlation with speech prominence [1]. (Energy is some times listed alone. We include it in the intensity category). 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? We will consider these questions below.

1) Syllable Duration: Syllable duration is an obvious feature. Speakers tend to stretch the constituent syllable durations when they try to emphasize a specific word [13]. Automatic syllable detection (ASD) has a long research history with an extensive literature on it. Early efforts to detect syllables were predominantly rule-based, and empirically derived, using various acoustic features extracted via signal processing approaches [36], [37]. Later research trends in locating syllable boundaries have moved from knowledge-based methods to data-driven approaches [9]. Here, a variety of speech features are extracted and models are built via statistical learning approaches such as HMM [38] or ANN (MLP) [8]. The major challenges are data size and the choice of the learning method to capture hidden patterns. Even with continuous improvements, neither approach provides satisfactory performance on continuous read speech, let alone spontaneous speech.

Due to the difficulties in locating syllable boundaries in the aforementioned methods, we have instead opted for using spectral-envelope-based speech rate information [17]. Such a method has apparent strengths: it does not rely on any statistical models that require considerable amounts of annotated data, such as HMMs in automatic speech recognition. Moreover, it does not rely on any explicit linguistic knowledge either. Hence, it could work in parallel with, or even as a front-end for, automatic speech recognition. Specifically, in this paper we adapt and expand on the algorithm proposed for speech rate estimation in [17]. We have shown such an approach could provide state of the art speech rate estimation performance [6]. As a side product, the method also provides syllable nucleus information. Further details are provided in Section III.

2) Pitch Patterns: Pitch patterns have long been believed to have strong correlation with prominence [3]. Again, in order to do efficient modeling, several excellent efforts have been made to quantitatively describe the complex pitch trajectory behavior. One trend attempts at categorizing 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 limitations in its self-completeness (and as a result many prosodic systems use a variant system modified from TOBI to meet specific needs [23]) and transcribing effectiveness. The other trend is to sacrifice or distort details of the pitch behavior and extract the very key desired features. Some widely used approaches in this context include the rise/fall/connection (RFC) model [12] and the tilt model [30]. Nevertheless, almost all of these methods treat pitch at the suprasegmental level. What we opt 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 be an adequate capture the pitch behavior.

The interesting problem is then to capture valid patterns from the pitch trajectory at this representational scale. Again, there has been a long tradition in investigating this problem. Even though the term "prominence" seldom appears in this literature, the related term "accent" is often referenced. Numerous hypotheses have been proposed as valid prominence or accent indicators and each claim validated using specific data and evaluation method. For example, the distance between F0 maxima and the corresponding virtual baseline at that timepoint has been proposed as a valid indication of accent [24]. Streefkerk [3] used pitch median and pitch range as a measure of accent. Sluijter and van Heuven [13] used pitch variation (pitch movement). Taylor [30] proposed to consider the accent characterized by a rise and followed by a fall. In the context of prominence analysis, such shape characterization may prove useful. Similarly, Tamburini [1] applied the sum of rise and fall amplitudes (directly measured from Taylor's [30] tilt parameter) for a more detailed pitch trajectory shape analysis. An interesting work by Knight [32] in fact shows that the pitch plateau is related to prominence perception. Also in that work, the absolute pitch level is shown to be an indicator of prominence.

While there are different hypotheses, and empirical findings, about the relation between pitch patterns and prominence that have been proposed in the literature, there is still no universally accepted model. There are even some arguments [22], though relatively few, saying that pitch plays relative little importance in the context of prominence. Assuming pitch-induced prominence only through a single f0 contour, Kochanski [31] recently showed that prominent and nonprominent syllables have similar f0 pattern histograms which in turn were used to indicate that they provide little or no discriminating ability.

Key questions that arise in this context relate to if and how these various findings relate with one another and if there exist rules/models that are more fundamental that can be learnt to adequately explain and reconcile all the aforementioned behavior. A major challenge to this of course is the availability of sufficient amounts of reliably transcribed data to facilitate modeling in statistically meaningful ways. Toward addressing this issue, we did a comparative analysis on various features to study their discriminative distance with respect to prominent and nonprominent class using Kullback–Leibler distance and through ANOVA analysis.

3) Spectral Intensity: Spectral intensity also correlates with prominence [1]. Prior research has shown that energy in the 500–2000 Hz band 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 two bands ([0–300] Hz, [2300–6000 Hz]), related to the nasal and fricative bands, have been shown to have not much acoustic correlation with prominence.

Beyond the straightforward measure of such subband energy, there has been research in measuring various transforms of spectral intensity. There has been a notion of "loudness" [25], an approximation to steady state perceptual loudness, with various measures for it such as through power spectral density [31]. We group all these measures into the spectral intensity measure since they are all highly correlated. In this paper, we primarily focus on the sonorant band, but instead of directly extracting the subband energy, we apply both a temporal and spectral correlation within that band. As a result vowels, due to their well-defined formant structure, could be boosted in the correlation envelope while the consonants and noises are not. In a preliminary study, we found that such an approach could help boost the center syllable magnitude and be more noise robust [6].

4) Relation Between Feature Categories: The aforementioned three feature categories come from three independently controllable aspects of speech production. From a signal-manipulation point of view, one can, for example, fix the syllable duration and change the pitch and intensity within that parameter range. Interestingly, however, these three features show strong correlation with one another under the condition of prominence.

Among the three categories of features, syllable duration and spectral intensity are straightforward to measure using well-established methods. However, as discussed in the previous section, capturing pitch pattern relations to prominence is challenging, and not adequately understood. We will report on a comparative study of all independent pitch behavior features in terms of their discriminative ability for prominence detection.

The remaining problem is how to fuse the information provided by the three features to optimize the prominence detection. One simple approach would be to assume no prior knowledge and feature independence, and set equal weights for all three features using the maximum entropy rule. However, a more reasonable, and often adopted, approach is to formulate an optimization problem by defining an evaluation function that can be derived using a development set. By maximizing the classification performance by proper tuning of the evaluation function, for example using EM algorithm, a more reasonable fusion scheme can be determined. The shortcoming of this approach is the potential risk that the evaluation function can be easily over-fit to the training set. Additionally, if we are tuning towards manually transcribed data sets, we should keep in mind the limitations of data size and transcriptions being not perfect to ensure statistical convergence. However, for POS-based evaluation, such limitations may be overcome at least partially due to the ability for processing large amounts of text data fairly reliably. In this paper, we will consider both supervised and unsupervised schemes. It should be noted that that the fusion we consider throughout paper includes both decision level and feature level combinations, and hence inherently accommodates multi-dimensional classifiers like support vector machines (Section V).

II. Data Description

Two different speech corpora were used for the studies described in this paper. The primary one is the ICSI-annotated switchboard data corpus used for analysis and testing of the methods proposed in this paper for unsupervised prominence detection based on POS correlation. A second, smaller corpus of spoken dialogs from role-playing scenarios which includes manually tagged word prominence information was used for a comparative analysis with the POS-based method.

A. ICSI Switchboard Data Corpus

In the Switchboard audio corpus two individuals discuss, over the telephone, a specific topic such as automobiles, sports, or politics for several minutes [18]. The data subset we used is from the phonetic switchboard transcription project at ICSI (University of California, Berkeley) that comprises 5682 speech spurts taken from the full Switchboard corpus [33]. It represents portions of 618 conversations from 750 speakers, of both genders, and spans a wide range of adult ages and dialectal patterns of American English. They were phonetically transcribed by a group of eight linguistics students all of whom had received previous training in phonetic transcription and general phonetics/phonology. Along with syllable transcription, word transcription and boundary timing are also provided. Manual prominence labels are not available.

B. SASO Dialog Corpus

To enable comparisons of the proposed POS-based method with prominence detection that uses manually tagged prominence labels, we use a dialog corpus that provides word level prominence markings made by human listeners. This helps us evaluate part-of-speech based prominence detection against direct prominence detection. The data are from role-playing dialogs between two individuals involved in negotiation and conflict resolution in military logistics as a part of the SASO project [20] at the University of Southern California. A subset of seven dialogs comprising 480 utterances was used for this study. Three native speakers of English, with speech processing/linguistics training, manually tagged these utterances. Instead of quantizing prominence into many different levels, the transcribers just marked the most prominent word(s). The

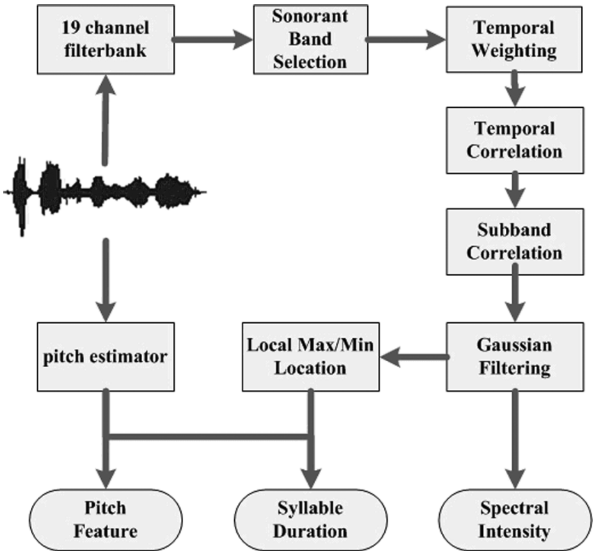


Fig. 1. Algorithm for locating syllable nucleus position from speech.

dataset has a total of 3247 words with 862 transcribed as prominent. Of the 862 768 were content words. The inter-transcriber agreement was 84%.

III. Algorithm Description

The proposed prominence detection relies on spectral and temporal features extracted from the speech signal to provide direct information about the syllable nucleus location, spectral intensity information as well as pitch pattern dynamics across an utterance. Details of each of these are given below.

A. *Syllable Nucleus Estimation*

Syllable nucleus estimation is a key ingredient of the algorithms we describe not only because syllable nucleus duration is an important acoustic correlate of prominence but it also provides the basis for defining other useful features. Locating a syllable using acoustic information is a difficult task, especially for spontaneous speech, and we describe a correlation-based signal analysis method to get a fairly robust estimate of the syllable nucleus.

The algorithm extracts the syllable nuclei from the correlation envelope of the speech signal. We extend the idea of subband correlation [17] and combine it with temporal correlation to obtain the final correlation envelope. The details of the algorithm, summarized in Fig. 1, are described below.

1) As described earlier in Section I-E, since the sonorant band is the most informative in the context of prominence detection, instead of performing a full band analysis, we only focus on the sonorant band. However, it should be noted that sonorant band also contains constants such as nasals and semivowels. This could introduce errors to the syllable nucleus detection. While such a problem is difficult to address in a general way, some of the post processing steps introduced in our approach, are aimed to handle, at least partially, this problem.
2) We make finer, and more, band divisions. They are Butterworth bandpass filters centered at 360|480|600|720|840|1000|1150|1300|1450|1600|1800|2000|2200 [19]. The purpose is to track formant movement by selecting the high-energy subbands for the correlation (refer to [6] for details).

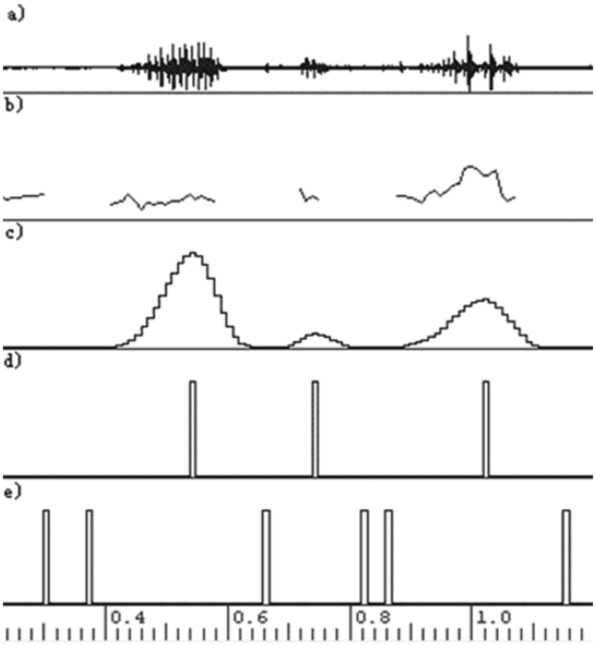


Fig. 2. Illustration of the results of the algorithm in Fig. 1. (a) Speech waveform. (b) Pitch trajectory. (c) Smoothed correlation envelope. (d) Location of maxima (peaks). (e) Location of minima (valleys).

3) Instead of doing a point-wise correlation, we do a selected subband correlation. Using experimental results from a development test set, we choose a subset of bands that have most energy and correlate them. Doing this allows the formant structure of vocalic regions to render the correlation envelope for vowels better emphasized compared to consonant and noise regions.
4) In order to make the syllable nucleus location more apparent and make the final correlation envelope smooth, not only is correlation performed spectrally, but also temporally. Here again, the size of the temporal window used for correlation was chosen empirically as a result of experiments on a development test (further details may be found in [6]).
5) To counter spurious peaks in the correlation envelope due to fricatives and other nonspeech noise, pitch verification is introduced to the algorithm. Peaks corresponding to unvoiced and nonspeech segments are rejected [for example, refer to Fig. 2(b)].
6) As additional measures to ensure robustness of syllable counting on the correlation envelope, further smoothing techniques are applied. One involves applying a Gaussian smoothing window, and the other is by setting a minimum threshold on peak heights. Again, these parameters are automatically learned using a development test.

As noted in the above description, the development test is utilized to tune the parameters of the various blocks of the algorithm. The goal is to maximize the correlation between the test syllable count and the transcribed count. Briefly, we first initialize all the parameters by Monte-Carlo simulation. The top performing candidate parameters in the development set are

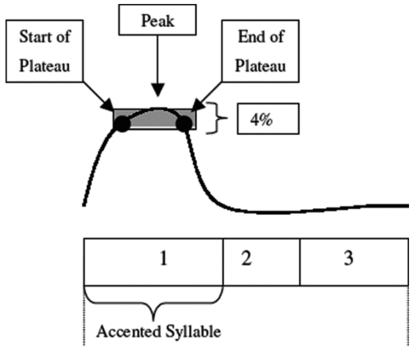


Fig. 3. Illustration of the plateau in the context of prominence, based on [30].

then chosen for a sensitivity analysis. This entailed perturbing each parameter by a small amount to seek an increase in performance until reaching a local maximum. Finally, the best parameter set is used to do the analysis.

Performing peak counting on the correlation envelope curves has been to shown to provide a good estimate of the true syllable counts, and has been advantageously used in estimating speech rate resulting in much improved performance on the ICSI switchboard data corpus. Preliminary experimental results on the Switch Board Corpus can be found in [6]. For example, the utterance in Fig. 2 has three syllables each centered in the peak of the curve. We will describe in the next section how we can get further useful features for prominence detection by exploiting the syllable estimation process.

B. Syllable Duration

We will describe in this section, how syllable duration and spectral intensity features could be derived from the correlation envelope. In addition to peak counting [Fig. 2(d)], we also keep the information of the valleys (local minima of the correlation envelope), as illustrated in Fig. 2(e). The duration score is retrieved by computing the valley-to-valley distance and normalizing it by the maximum duration. This we found was a more robust measure than peak-to-peak based distance: Since we apply pitch verification and thresholds for peak selection, there exist cases, albeit infrequently, that have no measurable peak between neighboring valleys. Hence, the distance we consider for duration calculation is between neighboring valleys that have a peak in between [Fig. 2(d) and (e)]. This method determines the syllable nucleus without requiring any linguistic model or statistical phonetic model and works as a real-time signal processing approach.

C. Spectral Intensity

The spectral intensity score is represented by the peak value normalized by the maximum peak. The peak directly comes from the selected subband correlation and the temporal correlation (the same used in Section III-A for correlation envelope generation). We however, for simplicity, do not consider the integration of the envelope [1] since this requires that syllable duration be recomputed.

D. Pitch Patterns

Pitch patterns are more difficult to characterize and quantify. We will consider various pitch pattern features.

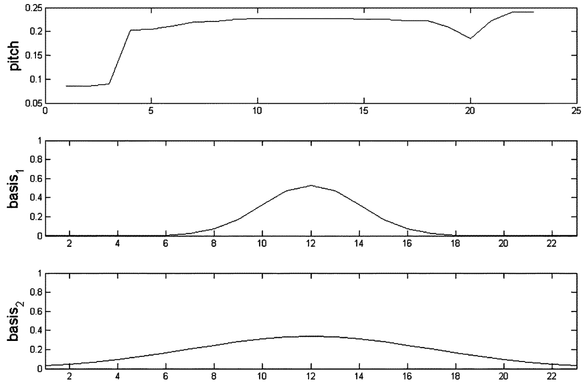


Fig. 4. Illustration of different basis windows to enable representing pitch plateau within-syllable.

1) Representation of Pitch Through Local Statistics: Here, we extract the following widely-used pitch measures in each syllable nucleus:

max	the maximum of pitch
min	the minimum of pitch
mean	the mean of pitch
median	the median of pitch
range	the range of pitch
std	the standard deviation of the pitch

2) Parametric Shape-Based Pitch Representations: We consider a more complex pitch characterization in this section. Taylor [30] proposed to characterize accent by a rise followed by fall in the pitch trajectory. Either manual labeling or automatic tagging could be used to identify such patterns. For automatic tagging process, a smoothing procedure followed by peak detection is proposed to characterize such pitch patterns [30]. The utility of such a representation of pitch in the context of prominence detection was demonstrated in a recent study by Knight [32] which shows that pitch plateaus are related to prominence perception. As illustrated in Fig. 3, they define a plateau as "being 4% down from any absolute peak in F0". The existence of such a pattern is deemed to be indicative of prominence.

The plateau could be viewed as an extension to Taylor's rise/fall model. However, smoothing followed by peak counting as originally proposed in [30] would not describe the plateau efficiently. Hence, to address this issue we apply a signal decomposition approach. Specifically, for representing pitch patterns within a syllable (Fig. 4, top row), we designed various Gaussian-shaped windows to serve as basis functions. By appropriate selection of the Gaussian window bases, we could model different plateau shape representations. Both the pitch and basis are normalized to have unit energy such that the inner product between normalized pitch and the different bases should be less or equal than one with being one implying that they have identical shape. The implementation is as follows. We first remove the syllable region without pitch. We take Gaussian windows with the same length of the leftover region. We choose five variations (0.2, 0.5, 1, 5, 10) that model the window width from narrow to wide. This will introduce five different inner products as five independent scores. It should be pointed out that the decomposition basis used here is Gaussian, and thus symmetric, and may not be optimal to capture the true pitch characteristics. However, as we will show in a later section that such a heuristic selection still is useful, at least for our problem of discriminating prominence.

E. Normalization

In the previous sections, we have introduced various acoustic features. In summary, we have one syllable duration feature, one spectrum intensity feature, six local pitch statistics features and five parametric shape-based features. Besides the last five parametric shape-based features that are inherently in the range of [0,1], all the other features are variable in range. Without losing generality, we normalize each feature category to a linear scale by the maximum value of the feature in the utterance. So each feature category are within the range of [0, 1]. This simplifies our later comparative study.

F. Prominence Score

Three types of features deemed useful for prominence detection were described along with how they could be derived directly from the acoustic speech signal. Each of these features is normalized at the utterance level such that they are mapped to float numbers in the range [0, 1]. Following the methodology outlined in Section I-E, for the first choice (supervised approach), we train the classifier with manually transcribed data. As the second option (unsupervised approach), equal weights are given to each feature type assuming that they arise from independent sources and no prior knowledge is available. Here, these features could thus be immediately fused to get a syllable level prominence score also in the range of [0, 1]. These two different prominence scores will be evaluated independently.

For most applications, prominence is evaluated at the word level. We derive the word-level prominence score from the syllable-level prominence score. Specifically, we use the most prominent syllable to represent the word prominence. This is based on the observation that the prominent syllable always carries most information on word prominence.

IV. Experiments

In Section I-E, we mentioned that the ideas of syllable duration and spectrum intensity are well established in terms of their correlation with prominence. However, on the other hand, the utility of the various pitch features that have been proposed in the context of prominence detection still needs to be addressed. In this section, we will study these issues using various experimental measures. For this purpose, we will use the SASO dialog corpus with manual prominence transcription. (See Section II).

A. Methodology

We apply the algorithm described in Section III and obtain the 13 normalized features for each word. All the words can be categorized into one of two classes: prominent and nonprominent. Then we collect the feature statistics for each class and study the discriminant distance for the probability density functions. Intuitively, a feature category that shows a large distance between these two classes indicates that it is a good feature for prominence detection.

B. Distance Measure

We apply the Kullback–Leibler distance and minimum classification error as measures of discrimination between prominent class and nonprominent class.

1) Kullback–Leibler Distance Measure: The Kullback–Leibler distance [15] is perhaps the most frequently used information-theoretic distance measure. If $p1$ and $p0$ are two probability densities each representing the pdf of prominent class and nonprominent class respectively, the Kullback-Leibler distance is defined to be

$$D(p1||p0) = \int p1(x) \log_2 \frac{p1(x)}{p0(x)} dx.$$

The baseline case is that the feature is uncorrelated with prominence. Thus $p0$ and $p1$ are similar and have a distance very close to 0. If the distance is larger, it means that this feature has a larger ability in discriminating between the prominent and nonprominent class.

2) Minimum Classification Error Measure: A larger KL distance implies a larger information divergence between prominent and nonprominent classes, but it might not necessarily mean we could have a a high classification rate by a simple decision threshold. One popular classification measure is by setting a decision threshold to minimize the classification error. The classification error (or correct) rate together with precision and recall rates are the typical measures of performance. These performance estimates are

$$E = \frac{fp + fn}{tp + fp + tn + fn} * 100\%$$
$$R = \frac{tp}{tp + fn} * 100\%$$
$$P = \frac{tp}{tp + fp} * 100\%$$

where fp and fn refers to false positive and negative and tp and tn are true positive and negative. The classification error E provides an overall error measure. The recall R value measures the percentage of correct predictions. The precision P gives the percentage of positives (prominent) that are correctly predicted. An important step in this measure is to balance the priors. In the particular situation of the SASO dialog corpus, the prominent and nonprominent classes are very unbalanced. For the discriminative study, we could simply assume equal priors and compare pdfs directly. For direct evaluation on the final fused prominence scores, we downsample the nonprominent class size to make it the same size as that of the prominent class.

C. Discriminative Analysis Results

In order to get a precise estimate, we estimate the pdfs through normalized histogram directly. The width of the histogram bin used was 0.01.

The overall results are given in Table I. To better interpret the results graphically, we use Parzen window generated pdfs in all the graphs in this section. We will next discuss the results for various feature categories.

1) Syllable Duration and Spectrum Intensity: We can observe (first two rows of Table I) that syllable duration and

TABLE I
OVERALL RESULTS

	KL-Distance (10^{-3})	Error Rate	Precision Rate	Recall Rate
Syllable duration	6.50	36.8	64.7	65.7
Spectrum Intensity	8.18	35.9	60.0	67.6
Local statistics:				
Pitch max	6.50	38.9	41.8	71.4
Pitch min	8.18	44.3	54.5	58.3
Pitch Median	6.38	40.3	46.8	65.9
Pitch Mean	6.13	40.1	45.9	66.5
Pitch Range	4.14	40.9	54.3	62.8
Pitch Std	4.14	42.9	59.0	60.1
Shape Features				
Sigma=0.2	4.24	42.6	92.8	55.9
Sigma=0.5	4.22	42.4	93.0	56.0
Sigma=1	3.91	42.2	93.7	56.0
Sigma=5	4.57	44.1	89.4	56.9
Sigma=10	5.12	55.9	70.5	58.0

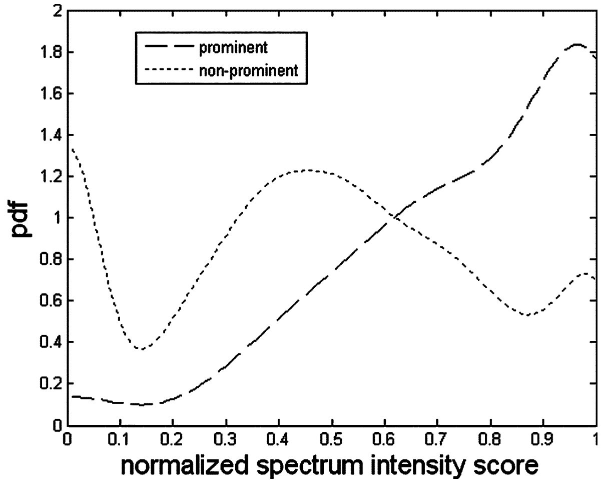


Fig. 5. Spectrum intensity distributions.

spectrum intensity are the top two useful features in discriminating prominence, with the results for spectrum intensity being slightly better than syllable duration. Figs. 5 and 6 show that they have similar pdf discrimination for the two classes under consideration.

2) Pitch Local Statistics: Table I shows that among all local statistics derived from pitch, pitch max provides the most discriminating information for prominence detection. However, the graphs of the pdfs (Fig. 7) do not depict this property clearly. This feature also has very low precision rate as in Table I. This in fact coincides with the results in [31]. We also find that pitch min, median, and mean all have similar pdf as pitch max which is illustrated in Fig. 7. In summary, we found that pitch local statistics only offer limited information about word prominence.

3) Parametric Shape-Based Features: The goal is to design a mechanism to capture the plateau event (see Section III-D). However, the result in Fig. 8 shows shows that the feature just by itself is not effective either from the KL-distance or classification point of view. However, we will demonstrate that pitch plateau is quite useful when it is combined with other features to do machine learning, demonstrating that it carries complementary information with respect to the other features.

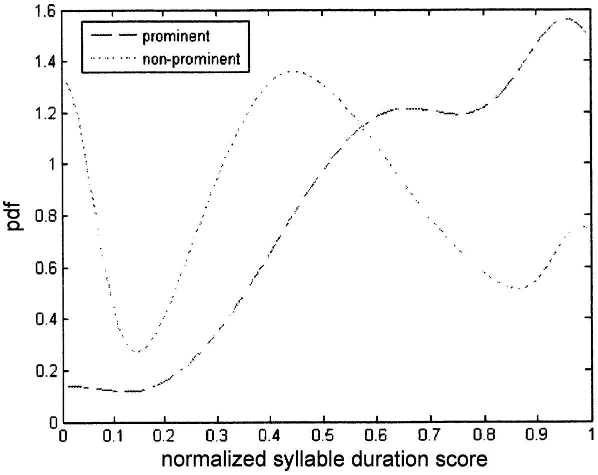


Fig. 6. Syllable duration distributions.

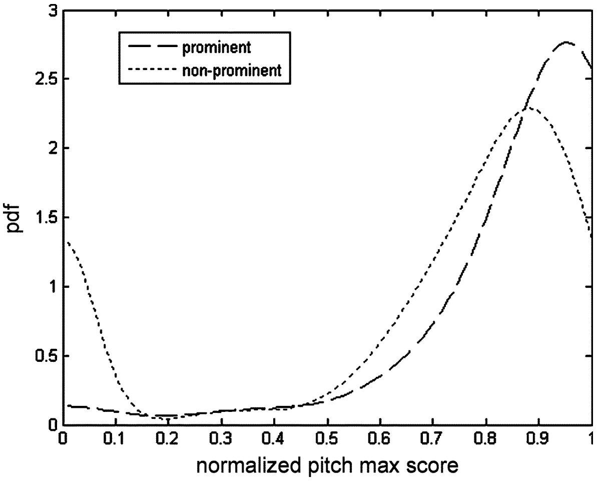


Fig. 7. Pitch Max. distributions.

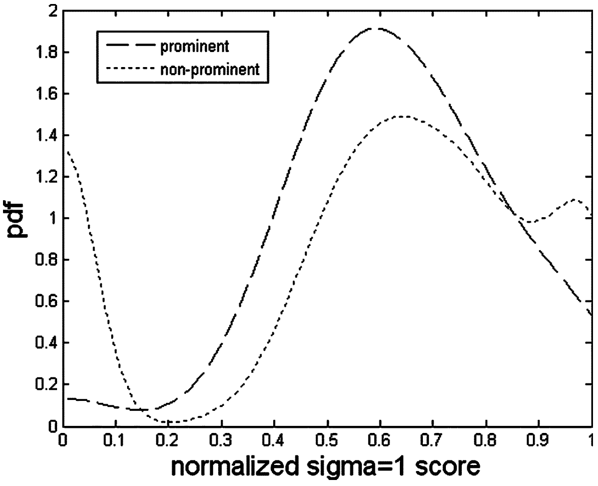


Fig. 8. Parametric shape-based feature with $\mathrm{sigma} = 1$.

TABLE II
OVERALL DISCRIMINANT RESULT WITH SUPERVISED FUSION

	Error Rate	Precision Rate	Recall Rate
C0	35.9	60.0	67.6
C2+C3	29.5	76.6	68.3
C1	26.6	78.7	71.1
C1+C2	26.2	79.7	71.3
C1+C3	25.4	79.2	72.6
C1+C2+C3	23.8	82.1	73.4

TABLE III
RESULTS WITH INDIRECT SUPERVISED FUSION

	Error Rate	Precision Rate	Recall Rate
Performance	27.3	66.9	89.7

V. FUSED PROMINENCE SCORE EVALUATION

In order to directly evaluate the utility of the score for prominence detection, we first performed experiments using the SASO corpus that had manually annotated word-level prominence tags. The data used in this experiment comprises role-play dialogs where human labelers tagged each word as being prominent or nonprominent (Section II).

A. Supervised Method

We fuse the various scores, computed as described in Section IV, to minimize the overall classification error. We make the following categorical notations:

C0	any single feature (any single row in Table I)
C1	syllable duration and spectrum intensity
C2	pitch local statistics
C3	parametric shape-based feature

We use support vector machine as our classifier. The "+" sign in Table II and the following discussion should be interpreted as a concatenation of the corresponding feature vector. The performance reported is using leave-one-out cross-validation (except C0).

We demonstrate in Table II the following.

- Syllable duration and spectrum intensity alone represents a reliable category of features for prominence detection. (C1)
- Parametric shape-based features demonstrate better performance than pitch local statistics when combined with C1. (C1 + C3 V.S. C1 + C2)
- Pitch features are useful in prominence detection although by themselves do not yield the best performance. The combination of all features provides the best performance. (C1 + C2 + C3)

B. "Indirect" Supervised Method for Fusion

In the next experiment, instead of relying on manually tagged prominence labels, we use the content word and function word information to derive the fusion weights. We adjust the weights of each feature category to maximize the divergence between these 2 classes. Results for prominence detection based on this method are given in Table III.

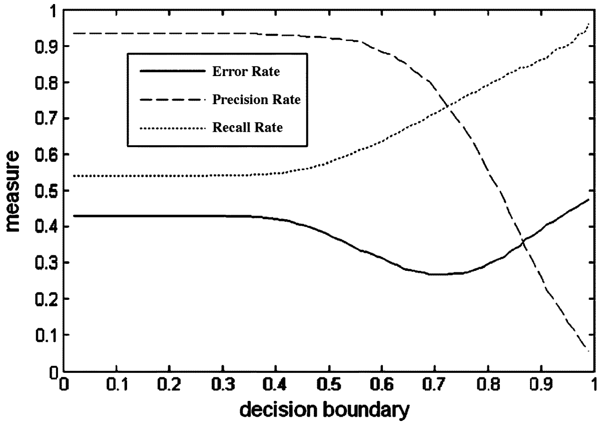


Fig. 9. Performance with unsupervised fusion.

TABLE IV
BEST PERFORMANCE WITH UNSUPERVISED FUSION

	Error Rate	Precision Rate	Recall Rate
Performance	26.6	80.0	70.0

We will see immediately see that this performance is somewhat worse than that obtained with the direct manual label information. This is not surprising since, although most prominent words tend to be content words, not all content words in an utterance tend to be prominent. Hence, the use of such information for supervised training may be somewhat misleading.

C. A Simple Unsupervised Method

The above supervised fusion method has the risk of over fitting the data. As an alternative, a simple unsupervised approach, that scores prominence with the simple average of syllable duration, spectrum intensity and pitch max (which is the best single feature from pitch feature category), can be considered. It should be noted that these features are not directly addable. Such operation is possible only when each feature category is normalized to be a score (in our case, a float number between 0 and 1).

Such operation could be formulated as

$$PS = (syl\_dur\_score + spec\_score + pitch\_\max\_score)/3.$$

The distributions of these scores are shown in Fig. 10. Results based on analysis of variance (ANOVA) indeed show that the scores for prominent and nonprominent cases are statistically different in a highly significant way, $\mathrm{p} < 1\mathrm{e} - 8$ (Fig. 11). In classification experiments, we obtain an error rate of 26.6% (refer to Fig. 9 and Table IV), which is below the best results by 2.8%.

VI. POS-BASED EVALUATION

In Section I-D, we discussed the idea of using POS as a means of indirectly evaluating the prominence detection algorithm. The underlying hypothesis is that words that appear more prominent in spoken language tend to belong to certain POS categories such as for instance those related to content words. This notion can be used to verify the prominence score derived from acoustic features. The evaluation process is summarized

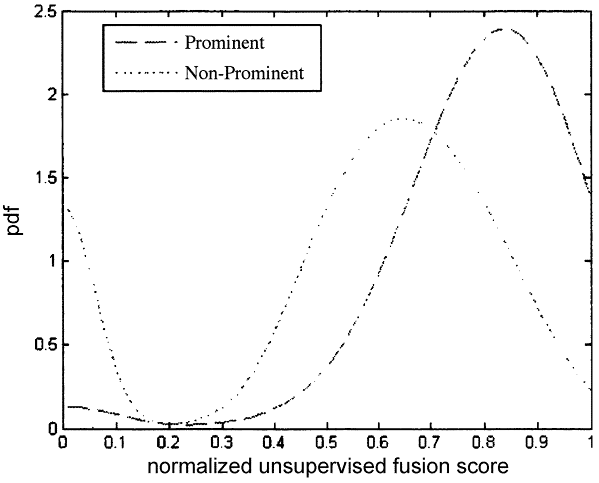


Fig. 10. Distribution of unsupervised prominence scores.

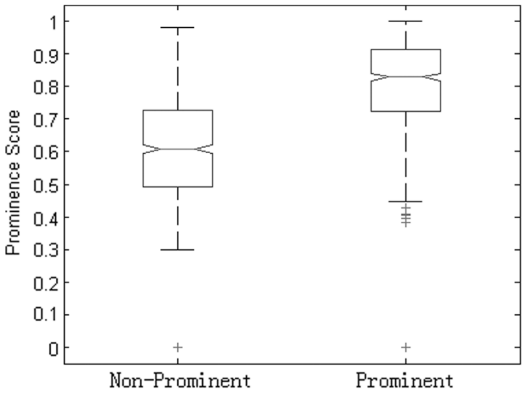


Fig. 11. Acoustic prominence scores for manually labeled prominent and non-prominent word categories in the SASO dialog corpus.

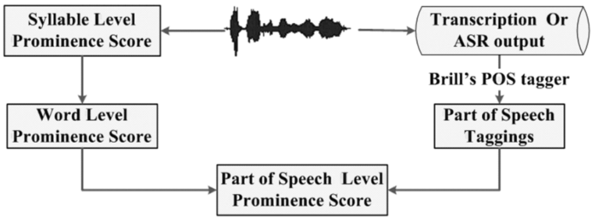

Syllable Level Prominence Score
Transcription Or ASR output
Brill's POS tagger
Word Level Prominence Score
Part of Speech Taggings
Part of Speech Level Prominence Score

Fig. 12. POS-based evaluation.

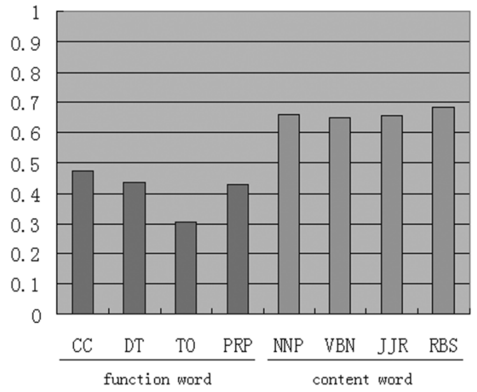


Fig. 13. Unsupervised prominence score for various POS categories in the Swithcboard corpus. CC: conjunction. DT: determiner. TO: infinitive marker. PRP: preposition. NNP: noun. VBN: verb. JJR: adjective. RBS: adverbs.

TABLE V
COMPARISON OF TWO FUSION METHODS

	Supervised score (10^{-3})	Unsupervised score (10^{-3})
KL-distance	4.58	2.93

in Fig. 12. Using the unsupervised method as outlined in the previous section (Section V-C), we can arrive at a word-level prominence score from speech immediately. For the supervised method (Section V-A), we start from the value of decision function of the support vector machine classifier and then normalize it to the range of [0, 1].

If we are given the true transcription or ASR output, the POS could be robustly retrieved through well-established statistical methods, e.g, Brill's tagger [16]. The error in this POS tagging process has been shown to be statistically insignificant.

The prominence scores for each POS category are then subject to further analysis. For both supervised and unsupervised methods, we obtained the prominence scores of the various POS categories in the ICSI Switchboard dataset. For illustration purposes, the results of the unsupervised case are shown in Fig. 13. (The supervised scores are similar but with a different scale.)

From Fig. 13, we could observe that there is a clear separation between the prominence scores for the content and function word classes. To investigate the statistical significance of this content word-function word score discrimination, we performed an ANOVA analysis. One way ANOVA showed that the prominence scores of content and function word classes are statistically distinct, $\mathrm{p} < 1\mathrm{e} - 6$. These results indicate that the prominence score has salient statistical divergence information between these two classes.

In Section V, the supervised method performs better than unsupervised method on the manually tagged prominence corpus. We wish to see if such fact could be extended to content/function word divergence. The KL-distances for both methods are in Table V.

The supervised prominence detection trained on manually tagged corpus derives a larger KL-distance between content and function word. This provides extra evidence for the validity of the supervised method, even though we train the supervised prominence detector on a small manually tagged dialog corpus.

While content/function word prominence difference in speech is a commonly accepted fact [30], evidence of which we have observed in our experiments, what we have evaluated so far is indeed not prominence directly but the indirect implication that the proposed acoustic score bears on prominence. However, the tags of content/function words could be retrieved through an automatic and objective process and the tagging error is almost negligible. This is apparently attractive than the manual subjective prominent/nonprominent tagging. So the statistical divergence information we observed between content and function word class could serve as an objective performance measure of prominence detection algorithm. Thus, different prominence detection algorithms could compare performance directly in this way without requiring laborious manual transcription.

Additionally, the prominence score we proposed in this paper could serve as an evidence of content word/function word distinction. Such a capability has potentially rich applications in automatic speech recognition and natural language understanding, although the experimental exploration of this topic is beyond the scope of this paper. Some potential applications are listed in the next section.

VII. Discussions and Conclusions

It is apparent from the results (Fig. 13) that content words have higher prominence scores than function words. There is a clear, statistically significant, distinction between the two classes. These results were computed from 5582 spontaneous speech utterances, and attest to statistical generality of the observation. Applying this algorithm to manually tagged prominence set data demonstrates even clearer discriminating ability (Fig. 11). In this sense, we can note that the proposed prominence score conveys useful information in the prosodic and POS realms in processing spontaneous speech.

The other issue about prominence detection is the role of context. The definition we quoted previously implies that prominence of a word should be measured by comparing with their surrounding context. Such information resides not only in the acoustic signal, the topic of this paper, but also in linguistic features [39]. The definition also implies that the comparison would be optimal locally, rather than globally. In this paper, we focus on just acoustic study ("ASR/transcription" free) and leave linguistic context analysis as future work. Additionally, we treat the acoustic comparison to be global. One justification to this is because our analysis considered mostly short utterances that makes global analysis a good approximation to local analysis. We also believe that the global method we derived would be easily ported to local analysis by defining the appropriate context region. Linguistic information could provide a further source to improve our algorithm in this regards.

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. Such applications are topics of ongoing investigations such as through the SASO project [20].

References

[1] F. Tamburini, "Automatic prosodic prominence detection in speech using acoustic features: an unsupervised system," in *Proc. Eurospeech 2003*, Geneva, Switzerland, pp. 129–132.

[2] B. M. Streefkerk, "Acoustical correlates of prominence: A design for research," in *Proc. Inst. Phon. Sciences*, 1997, vol. 20.

[3] B. M. Streefkerk, L. C. W. Pols, and L. F. M. ten Bosch, "Acoustical features as predictors for prominence in read aloud Dutch sentences used in ANN's," in *Proc. Eurospeech '99*, Budapest, 1999, pp. 551–554.

[4] B. Pickering, B. Williams, and G. Knowles, "Analysis of transcriber differences in SEC," in *Working With Speech*, G. Knowles, A. Wichmann, and P. Alderson, Eds. London, U.K.: Longman, 1996, pp. 61–86.

[5] J. Terken, "Fundamental frequency and perceived prominence of accented syllables," *J. Acoust. Soc. Amer.*, vol. 95, no. 6, pp. 3662–3665, Jun. 1994.

[6] S. Narayanan and D. Wang, "Speech rate estimation via temporal correlation and selected subband correlation," in *Proc. ICASSP*, Philadelphia, PA, Mar. 2005.

[7] T. Portele and B. Heuft, "Towards a prominence-based synthesis system," *Speech Commun.*, vol. 21, pp. 61–71, 1997.

[8] S.-L. Wu, M. L. Shire, S. Greenberg, and N. Morgan, "Integrating syllable boundary information into speech recognition," in *Proc. ICASSP 97*, vol. 2, pp. 987–990.

[9] A. W. Howitt, "Automatic Syllable Detection for Vowel Landmarks," Ph.D. dissertation, MIT, Cambridge, MA, 2000.

[10] E. Campione and J. Veronis, "A multilingual prosodic database," in *Proc. ICSLP98*, Sydney, Australia, 1998.

[11] J. Pierrehumbert and J. Hirschberg, "TOBI: A standard for labeling English prosody," in *Proc. ICSLP*, 1992.

[12] P. A. Taylor, "The rise/fall/connection model of intonation," *Speech Commun.*, vol. 15, pp. 169–186, 1995.

[13] A. Sluijter and V. van Heuven, "Acoustic correlates of linguistic stress and accent in Dutch and American English," in *Proc. ICSLP96*, Philadelphia, PA, 1996, pp. 630–633.

[14] V. Strom, "Detection of accents, phrase boundaries, and sentence modality in German with prosodic features," in *Proc. Eurospeech95*, Madrid, Spain, vol. 3, pp. 2039–2041.

[15] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Stat.*, vol. 22, pp. 79–86, 1951.

[16] E. Brill, "A report of recent progress in transformation-based error-driven learning," in *AAAI*, 1994.

[17] N. Morgan and E. Fosler-Lussier, "Combining multiple estimators of speaking rate," in *Proc. IEEE ICASSP'98*, Seattle, WA, May 1998.

[18] S. Greenberg, "The switchboard transcription project," F. Jelinek, Ed., 1996 Large Vocabulary Continuous Speech Recognition Summer Research Workshop Technical Reports ch. 6, Center for Language and Speech Processing, Johns Hopkins Univ. Baltimore, MD, Apr. 1997, Research Notes No. 24.

[19] Speech Filing System [Online]. Available: http://www.phon.ucl.ac.uk/resource/sfs

[20] D. Traum, W. Swartout, J. Gratch, S. Marsella, P. Kenny, E. Hovy, S. Narayanan, E. Fast, B. Martinovsky, R. Baghat, S. Robinson, A. Marshall, D. Wang, S. Gandhe, and A. Leuski, "Virtual humans for non-team interaction training," in *6th SIGdial Workshop on Discourse and Dialogue*, Lisbon, Portugal, Sep. 2–3, 2005.

[21] M. E. Beckman and J. J. Venditti, "Tagging prosody and discourse structure in elicited spontaneous speech," in *Proc. Science and Technology Agency Priority Program Symp. on Spontaneous Speech*, Tokyo, Japan, 2000, pp. 87–98.

[22] R. Silipo and S. Greenberg, "Prosodic stress revisited: Reassessing the role of fundamental frequency," in *Proc. NIST Speech Transcription Workshop*, May 2000.

[23] C. W. Wightman, "ToBI or not ToBI?," in *Speech Prosody 2002*, France, Apr. 11–13, 2002.

[24] J. M. B. Terken, "Variation of accent prominence within the phrase: models and spontaneous speech data," in *Computing Prosody for Spontaneous Speech*, Y. Sagisaka, W. Campbell, and N. Higuchi, Eds. Berlin, Germany: Springer-Verlag, 1997, pp. 95–116.

[25] H. Fletcher and W. A. Munson, "Loudness, its definition, measurement, and calculation," *J. Acoust. Soc. Amer.*, vol. 5, pp. 82–108, 1933.

[26] J. Pierrehumbert*, The Phonology and Phonetics of English Intonation*. Cambridge, MA: MIT Press, 1980.

[27] J. Hirschberg, "Pitch accent in context: predicting intonational prominence from text," *Artific. Intell.*, vol. 63, no. 1–2, pp. 305–340, 1993.

[28] C. Gussenhoven and A. C. M. Rietveld, "Fundamental frequency declination in Dutch: testing three hypothesis," *J. Phonetics*, vol. 16, pp. 355–369, 1988.

[29] C. Widera, T. Portele, and M. Wolters, "Prediction of word prominence," in *Eurospeech*, 1997, pp. 999–1002.

[30] P. Taylor, "Analysis and synthesis of intonation using the tilt model," *J. Acoust. Soc. Amer.*, vol. 107, no. 3, pp. 1697–1714, 2000.

[31] G. Kochanski, E. Grabe, J. Coleman, and B. Rosner, "Loudness predicts prominence: fundamental frequency lends little," *J. Acoust. Soc. Amer.*, vol. 118, no. 2, pp. 1038–1054, Aug. 2005.

[32] R.-A. Knight, "The realisation of intonational plateaux: Effects of foot structure," in *Cambridge Occasional Papers in Linguistics*, L. Astruc and M. Richards, Eds. Cambridge, U.K.: Cambridge Univ. Press, 2004, vol. 1, pp. 157–164.

[33] S. Greenberg, J. Hollenback, and D. Ellis, "Insights into spoken language gleaned from phonetic transcription of the switchboard corpus," in *Proc. Int. Conf. Spoken Language Processing*, Philadelphia, PA, 1996.

[34] F. Tamburini and C. Caini, "An automatic system for detecting prosodic prominence in American English continuous speech," *Int. J. Speech Technol.*, vol. 8, pp. 33–44, 2005.

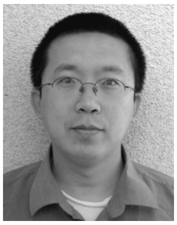
[35] P. C. Bagshaw, "An investigation of acoustic events related to sentential stress and accents, in English," *Speech Commun.*, vol. 13, pp. 333–342, 1993.

[36] D. J. Hermes, "Vowel onset detection," *J. Acoust. Soc. Amer.*, vol. 87, no. 2, pp. 866–873, Feb. 1990.

[37] N. N. Bitar and C. Y. Espy-Wilson, "The design of acoustic parameters for speaker independent speech recognition," in *Eurospeech*, Rhodes, Greece, Sep. 1997, pp. 1239–1242.

[38] P. D. Green, N. R. Kew, and D. A. Miller, "Speech representations in the SYLK recognition project," in *Vis. Represent. Speech Signals*, M. Cook, S. Beet, and M. Crawford, Eds. New York: Wiley, 1993.

[39] J. M. Brenier, D. Cer, and D. Jurafsky, "The detection of emphatic words using acoustic and lexical features," in *Proc. EUROSPEECH-05*, 2005.



Dagen Wang received the B.S. and M.S degrees in from the Electrical Engineering Department, Peking University, Beijing, China, in 1997 and 2000 and the Ph.D. degree in 2006 from the Electrical Engineering Department, University of Southern California, Los Angeles (USC).

He was with Intel China Research Center, Beijing, first as a Research Engineer in the Speech Group and later as a Researcher in the Programming Systems Group. His research interests are in the areas of spontaneous speech processing, including prosodic modeling and joint acoustic/language modeling, and signal processing and artificial intelligence with applications to speech, language, and human computer interaction problems.

Shrikanth Narayanan (SM'02) received the Ph.D. degree from the University of Southern California, Los Angeles (USC), in 1995.

He was with AT&T Research (originally AT&T Bell Labs), first as a Senior Member, and later as a Principal Member, of its Technical Staff from 1995–2000. Currently, he is a Professor of Electrical Engineering, Linguistics and Computer Science at USC. He is also a member of the Signal and Image Processing Institute and a Research Area Director of the Integrated Media Systems Center, a National Science Foundation (NSF) Engineering Research Center, at USC. His research interests include signals and systems modeling with applications to speech, language, multimodal, and biomedical problems. He has published over 175 papers and has ten granted/pending U.S. patents.

Dr. Narayanan was an Associate Editor of the IEEE TRANSACTIONS OF SPEECH AND AUDIO PROCESSING (2000–2004) and is currently an Associate Editor of the *IEEE Signal Processing Magazine*. He serves on the Speech Processing and Multimedia Signal Processing technical committees of the IEEE Signal Processing (SP) Society and the Speech Communication committee of the Acoustical Society of America. He is a Fellow of the Acoustical Society of America, a Senior Member of IEEE, and a member of Tau-Beta-Pi, Phi Kappa Phi, and Eta-Kappa-Nu. He is the recipient of the NSF CAREER award, USC Engineering Junior Research Award, USC Electrical Engineering Northrop Grumman Research Award, a Provost fellowship from the USC Center for Interdisciplinary Research, a Mellon Award for Excellence in Mentoring, and a co-recipient of a 2005 Best Paper Award from the IEEE SP Society.