Semantic Abnormality and its Realization in Spoken Language

Shimei Pan †

IBM Watson Research Center
30 Saw Mill River Road
Hawthorne, NY 10532
shimei@us.ibm.com

Kathleen McKeown

Computer Science Department
Columbia University
New York, NY 10027
kathy@cs.columbia.edu

Julia Hirschberg

AT&T Labs-Research
180 Park Ave
Florham Park, NJ 07932
julia@research.att.com

Abstract

In this paper we investigate the relationship between various lexical and prosodic features and *semantic abnormality*, the occurrence of unusual or unexpected events, in generating speech for MAGIC, which employs a Concept-to-Speech system to generate post-operative reports for patients who have undergone bypass surgery. Using the speech corpus collected for this application, we conducted empirical analysis to systematically discover significantly correlated prosodic and lexical features. The automatically learned abnormality model not only can be used in building comprehensive prosody prediction systems for Concept-to-Speech generation, but also help identify unusual information during speech analysis and understanding.

1. Introduction

Assigning prosody in speech generation systems is critical to system performance: not only should one convey meanings naturally, as humans would, but also one should convey it effectively. Previously, we have explored the role of various syntactic, semantic and lexical features in producing natural prosodic variations [1, 2, 3]. For example, in [2], we found that in addition to part-of-speech, word informativeness is a good predictor of pitch accent placement. We have also empirically verified the usefulness of word predictability in noun accent prediction [3]. Most features that we have investigated so far have been surface features that can be obtained automatically through text analysis. For example, part-of-speech can be obtained from free texts through part-of-speech tagging, while word predictability and informativeness are statistically modeled using text corpora. Thus, they are useful for both Concept-to-Speech (CTS) and Text-to-Speech (TTS) systems. In contrast to surface features, deep semantic features are usually much harder to explore because they are much harder to parse in free text. As a result, currently, they are primarily useful for Concept-to-Speech systems. Typical deep semantic features include information importance, semantic concepts and roles, semantic categoreis and intended speaker goal (e.g., convey abnormality, convey urgency). Since most automatic prosody modeling research has been conducted in Text-to-Speech research, the influence of deep semantic features in prosody modeling has not been extensively studied and empirically verified. In this paper, we focus on one such feature, semantic abnormality. We want to empirically analyze how it is communicated in speech and how it can be used in CTS generation.

† This work was conducted while the author was in Columbia University and was supported by NSF Grant IRI9528998, NLM Grant R01 LM06593-01 and Columbia University Center for Advanced Technology.

The prosody modeling work discussed here is part of a larger effort in developing an intelligent multimedia presentation generation system called MAGIC (Medical Abstract Generation for Intensive Care) [4]. In MAGIC, given a patient's medical record stored at Columbia Presbyterian Medical Center (CPMC)'s databases, the system automatically generates a post-operative status report for a patient who has just undergone bypass surgery. There are two media-specific generators in MAGIC: a graphics generator that automatically produces graphical presentations from database entities, and a CTS generator that automatically produces coherent spoken language presentations from these entities. The graphical and the speech generators communicate with each other on the fly to ensure a coordinated presentation.

In the rest of the paper, we will define semantic abnormality, describe the prosodic features we investigated as possible correlates of it, and describe our experiments. To explain some somewhat puzzling results we encountered in our analysis, we introduce a new feature, break index difference, which proved to be a useful indicator of semantic abnormality. In addition, we also show how *lexical unexpectedness* is related to abnormality and how to use regression models to systematically combine various prosodic and lexical features in predicting abnormality.

2. Definition of semantic abnormality

Semantic abnormality refers to something which does not usually occur in a particular context. So, it is a measure of unexpectedness. Abnormality defined in this way is domain and context dependent. For example, for the general population, blood pressure of 170/100 is considered high and therefore, abnormal. However, for patients who need a cardiac surgery, this value may still be expected, and is treated as normal. Thus, if a patient's condition is unexpectedly good or bad, both are categorized as abnormal. Identifying abnormality in general is not trivial. In our application, because we apply it in a specific domain, the identification task can be done reliably by a domain expert.

3. Speech corpus

To empirically investigate how semantic abnormality is related to prosody, we collected a speech corpus in Columbia Presbyterian Medial Center (CPMC) Cardiac Intensive Care Unit (ICU), where an Operation Room (OR) doctor informed residents, nurses and doctors in the ICU about the post-operative status of a patient. The main purpose of the speech was to inform ICU doctors and nurses about what happened to the patient before during and after a cardiac operation. The corpus primarily contains spontaneous monologues from multiple speakers. In addition, the speech style in general is calm and unemotional. Thus, unlike most public speakers or news anchors, the prosodic cues used by the physicians were subtle. In addition, the collected speech was transcribed orthographically by a medical professional and then prosodically labeled by a ToBI (Tone and Break Index) expert using the ToBI intonation labeling convention [5]. Here are the main ToBI features labeled: *pitch accent*, *break index*, *phrase accent*, *boundary tone*, and *HiF0*.

Moreover, we also annotated the speech corpus with semantic abnormality. Based on the ToBI break indices, utterances in the corpus were first separated into intermediate (minor) phrases. As a result, the basic units in this study are intermediate phrases. We also asked a doctor to categorize whether the information conveyed in each intermediate phrase is abnormal (1) or not (0). Sometimes, the doctor assigned a single tag to several adjacent phrases because each phrase by itself did not contain enough information for her to make a judgment. The final corpus includes eight speech segments and contains 784 intermediate phrases, 114 of which are categorized as abnormal, and all of which are annotated with both abnormality and ToBI prosodic features.

4. Prosodic correlates

Based on informal observations, semantic abnormality may be associated with a combination of prosodic features that appear to be intended to draw the listener's attention to the information being conveyed. For example, speaking rate, pitch range and F0 changes can all be used to highlight information in speech. A speaker may increase or decrease her speaking rate. In addition, expanded pitch range, increase in loudness, and increase in the number of accented items, and more frequent pauses often appeared to be associated with information the speaker wished to make more prominent. So, these features were all the candidates for our investigation.

Overall, we explored seven prosodic features: *speaking rate*, *HiF0*, *RMS total*, *F0 total*, *break index before*, *break index after*, and *accent probability*. All of them are computed for each intermediate phrase. For example, *Speaking rate* is defined as the number of syllables per second in an intermediate phrase. It is computed semi-automatically. First, the number of syllables in a word is extracted from a manually constructed lexicon. Then a script is used to automatically compute the average speaking rate for each intermediate phrase. *HiF0* is a general measure of a speaker's pitch range. Instead of directly using the F0 maximum, we use the manually labeled *HiF0* in ToBI because of its robustness. The next two measures, *RMS total* and *F0 total*, are automatically computed. First, we extracted both RMS and F0 from speech files using the *XWAVES* toolkit. Since they are not directly associated with an intermedia phrase, we use the sum of RMS and F0 over each intermediate phrase instead. The next feature, *Break index*, is a major ToBI feature. It is an indication of the relative level of juncture between orthographic words, acoustically signaled by a combination of F0, duration and optional pauses. Here, we investigate both the *break index* before and after an intermediate phrase. Since the smallest units are intermediate phrases, their values can only be "3" or "4". Finally, *Accent probability* is defined as the percentage of words that are accented in an intermediate phrase. Intuitively, we expect that low speaking rate, high HiF0, high RMS and F0 total, larger break index before and after a phrase, and high accent probability may signal abnormality. To reduce the influence of inter-speaker variations, we normalized *speaking rate*, *HiF0*, *RMS total* and *F0 total* before conducting our empirical analysis.

5. Correlation test

To understand how prosodic features are associated with abnormality, we performed a set of correlation analyses based on Spearman's rank-based correlation test. The test results shown in Table 1 reveal two types of information: the correlation coefficient *rho* and its associated statistical significance *p-value*.

Table 1: *Abnormality and Prosody.*

Prosodic Features	Rho	P-value
Speaking Rate	0.02	0.60
HiF0	0.13	< 0.01
RMS total	0.0022	0.96
F0 total	0.12	< 0.01
Break Index Before	-0.08	0.05
Break Index After	0.086	0.04
Accent Probability	-0.04	0.32

The test results demonstrate that *HiF0*, *F0 total*, *Break Index Before*, and *Break Index After* are significantly correlated with abnormality with $p - value <= 0.05$. Since their correlation coefficients are positive, higher *HiF0*, higher *F0 total* and more significant break index afterwards are more likely to be associated with abnormal information ($rho > 0$). However, for *break index before*, although it shows a certain degree of correlation, the association is negative ($rho < 0$), which means the *break index* is less significant before phrases containing abnormal information. This is inconsistent with our intuition. In general, significant prosodic phrase boundaries are associated with important information. To explain the negative correlation between *break index* and semantic abnormality, we conducted additional experiments with a new feature, *break index difference*. Our analysis results indicate a significant positive correlation between the abnormality and the *break index difference* before an intermediate phrase.

6. Break index difference and semantic abnormality

After analyzing our corpus, we speculate that the negative correlation is a result of using break index in simultaneously conveying several kinds of information, such as semantic importance, information structure, and semantic/syntactic structure. In our corpus, many sentences follow the following pattern: *theme* + *rheme*. *Theme* is the current topic as well as the connector to a previous context. In contrast, *rheme* communicates new information about a *theme*. Thus, based on our definition of abnormality, *rhemes* may be considered more important. Here is an utterance from our corpus: "(He is uh) (a heavy alcohol drinker)". "He is uh" is the *theme*, and "a heavy alcohol drinker" is the *rheme*. Prosodically, the utterance consists of two intermediate phrases: "He is uh" and "a heavy alcohol drinker". Since the boundary before "He is uh" is a sentence boundary, its break index is almost always "4". While the break index before "a heavy alcohol drinker" usually is less significant (can be either "3" or "4"). In term of semantic abnormality, however, the first phrase was labeled as normal and the second one was labeled as abnormal. Thus, there exists a mild negative correlation between abnormality and the *break index*.

Moreover, the strength of a break index is also affected by an utterance's semantic and syntactic structure. For example, a sentence boundary or a clause boundary is often signaled by a significant prosodic phrase boundary. In contrast, the boundary between an article and a head noun as in the phrase "a patient" usually is insignificant. After analyzing the corpus, we realized that significant prosodic phrase boundaries that are not licensed by the utterance's syntactic/semantic structure can signal abnormality. Here is another example from our corpus, "(She was uh) (re-admitted) (on ten twenty two) (with) (staph aureus sepsis)". This sentence contains five intermediate phrases. The second phrase "re-admitted" and the fifth one "staph aureus sepsis" were labeled as abnormal. In both cases, the prosodic phrase boundaries should be insignificant in normal speech because in the first case, it is between an auxiliary verb and a main verb within a verb phrase, while the second one is between a preposition and a noun phrase within a preposition phrase.

To verify that it is not the absolute break index that is important in conveying abnormality, but the difference between the break index observed and the break index that is licensed by its associated semantic/syntactic structure, we introduce a new feature called *break index difference*. We want to test whether this new feature is significantly and positively associated with abnormality.

Before we computed the new feature, we need to compute an index that measures the significance of a semantic/syntactic constituent boundary. The semantic/syntactic structure of a sentence used here was based on systemic grammar [6]. In systemic grammar, the process (ultimately realized as the verb) is the core of a clause's semantic structure. Obligatory roles, called participants, are associated with each process. Usually, participants convey who/what is involved in the process. The process also has peripheral roles called circumstances. Circumstances answer questions such as when/where/why. Given such a structure, it is quite straightforward to define the semantic/syntactic constituent boundaries. For example, in a sentence like "Her hospital course was complicated by respiratory failure requiring nitric oxide and mechanical ventilation", the boundary before *her* is a sentence boundary (SB) and that between *failure* and *requiring* is a clause boundary (CB). Similarly, between *course* and *was*, there is a participant boundary (ParB) and between *complicated* and *by* there is a circumstance boundary (CirB), The boundary between *nitric* and *oxide* is a word boundary (WB). We also heuristically defined the order among them. For example *SB* is more significant than *CB*, and *CB* is more significant than *CirB*, etc. Finally, these boundaries were also mapped to a number from 1 to 4. For a detail explanation of the sentence structure and the semantic/syntactic constituent boundaries, check [1].

Before proceeding with our statistical analysis, we manually cleaned all the utterances in the corpus. All the disfluencies and repairs were removed and the clean corpus contained only grammatical sentences or sentence segments. Then we assigned a semantic/syntactic boundary to all locations between two adjacent words. Finally the difference between the break index used by the speaker and its semantic/syntactic boundary index was computed. Based on this information, we investigated two new variables: the boundary difference before and after an intermediate phrase. Table 2 shows the results of the correlation tests.

As we expected, the boundary difference before a phrase is significantly associated with abnormality. The larger the difference is, the more likely it will be followed by a piece of abnormal information. The other new feature, the boundary difference after a phrase, however, does not seem to signal abnormality.

Table 2: *Abnormality and Index Difference.*

Prosodic Features	Rho	P-value
Boundary Difference Before	0.125	< 0.01
Boundary Difference After	-0.04	0.29

7. Lexical unexpectedness

Our previous analysis demonstrates that semantic abnormality is associated with a set of prosodic features. In addition, abnormality can also be communicated lexically. Therefore, certain lexical properties may also be correlated with semantic abnormality. For example, rare words may convey rare concepts. Since our semantic abnormality is defined through semantic unexpectedness, unexpected words may also signal abnormal situation. As a result, the next candidate to investigate in our abnormality modeling is a word's lexical unexpectedness.

In a separate study [2], we define a word unexpectedness as the negative log of the probability of seeing a word in a corpus. According to formula 1, the unexpectedness of a common word will be low because its occurrence is high. It should be high for rare words because its occurrence in the corpus will be low.

$$Unexpectedness(W_i) = -\log \frac{Freq(W_i)}{\sum_{i=1}^{n} W_i} \qquad (1)$$

Where n is the number of unique words in a corpus and $Freq(W_i)$ is the occurrence of word W_i in the corpus.

In that study, we used a much larger text corpus which contains over 7000 discharge summarizes for patients who also underwent surgery in CPMC. Since the majority of the patients underwent cardiac surgery, the text and speech corpus contains similar content. Using a similar corpus ensures the accuracy of the word unexpectedness metric because of its domain dependency. For example, the word "finance" is unexpected in the cardiac intensive care domain, thus its unexpectedness is high. In contrast, in a Wall Street Journal corpus, it is a common word and its unexpectedness is low.

In order to compute the correlation between semantic abnormality and lexical abnormality, we first calculated the average lexical unexpectedness for all the words within an intermediate phrase, the basic unit in our analysis. Then we applied the same correlation test. As we expected, the average lexical unexpectedness is significantly associated with semantic abnormality with $rho = 0.15$ and $p - value < 0.01$.

In addition, unexpected words are also associated with many prosodic parameters. Based on our results shown in table 3, unexpected phrases are spoken more rapidly ($p-value < 0.01$). They are more likely to follow larger boundary differences ($p - value < 0.01$) and to be followed by larger break indexes ($p - value < 0.01$). In addition, words within such phrases are more likely to be accented ($p - value < 0.01$). However, the break index differences after them tend to be smaller.

8. Modeling abnormality in spoken language

Our analyses have shown that, in our domain, semantic abnormality is reliably associated with a set of prosodic and lexical features. These analyses, however, did not take the possible interactions among different prosodic features into consideration.

Table 3: *Lexical Unexpectedness and Prosody.*

Prosodic Features	Rho	P-value
Speaking Rate	0.15	< 0.01
HiF0	-0.003	0.94
RMS total	0.048	0.24
F0 total	0.046	0.26
Break Index Before	-0.02	0.60
Break Index After	0.11	< 0.01
Accent Probability	0.27	< 0.01
Break Index Difference before	0.30	< 0.01
Break Index Difference after	-0.11	< 0.01

Next, we demonstrate how a combination of all these features is related to abnormality.

We employ the generalized linear regression model to represent the relation between abnormality and all the features. Unlike the traditional linear model, the generalized linear model employs separate link functions to allow for nonlinearity. Since the semantic abnormality feature takes one of two values, the logistic regression model is specially designed for modeling binary and more generally, binomial data. This model can then be fitted by iteratively re-weighted least squares. We used the *Splus* statistical package [8] for the analysis. The data were analyzed in a step-wise fashion. Initially, all the prosodic and lexical features, no matter whether they were correlated with abnormality or not during the individual correlation analysis, were included. At each step, a single feature was selected and dropped based on how well the new model fit the data. The dropped features were either irrelevant or redundant. Formula 2 shows the final abnormality model learned from this analysis:

$$\begin{aligned} Abn &= 0.62 * HiF0 + 0.81 * F0_t \\ &\quad -0.55 * Index_p + 0.25 * IndexDiff_p \\ &\quad +0.46 * UnExpect - 0.75 * Prob_{ac} \\ &\quad -2.69 \end{aligned} \tag{2}$$

where *Abn* is abnormality; $F0_t$ is the *total F0*; $Index_p$ is the break index before the phrase; $IndexDiff_p$ is the break index difference before the phrase; $UnExpect$ is the average word unexpectedness and $Prob_{ac}$ is the percentage of accented words in an intermediate phrase.

Based on the combined model, $HiF0$, $F0_t$, $IndexDiff_p$, $UnExpect$ positively influence abnormality. The larger those values are, the more likely the conveyed information is abnormal. In contrast, $Index_p$ and $Prob_{ac}$ negatively contribute to abnormality. These observations are quite consistent with our individual association analysis.

9. Applications

As part of an ongoing effort, we are working on applying the derived results in MAGIC CTS generation. For example, since the break index difference is significantly correlated with abnormality, we may adjust the break index before a phrase so that abnormal information can be communicated more effectively. In addition, we can also use the learned model in speech perception and understanding. Basically, the learned regression model may directly serve as a classifier to identify abnormal information in spoken utterances, using both prosodic and lexical cues.

10. Related Work

This study is related to earlier work on analyzing and producing affective (emotional) prosodic patterns [9, 10]. For example, one of the affects modeled in [9] is *surprise*. Cahn used a combination of F0 parameters, such as pitch range, and accent shape, timing parameters, such as pauses, and speaking rate, and voice quality parameters, such as breathiness and loudness, to construct a *surprise* production model. Subjective evaluation demonstrates the promise of this model. In addition, emphatic speech patterns, such as contrastive accent patterns, also involve similar speech parameters. In [11], Prevost used both pitch amplitude and different types of pitch accent to realize contrastive accent in synthesized speech.

11. Conclusions

Due to the availability of deep semantic information, so far, empirically verifying how deep semantic information is communicated in spoken language is still quite new. In this paper we have investigated how prosodic and lexical cues are used to communicate semantic abnormality. Among all the features tested, HiF0, F0 total, break index after, break index difference before and word unexpectedness are significantly correlated with abnormality in both correlation and regression tests. In addition, accent probability is chosen by the combined abnormality model as a useful feature. Other features, such as speaking rate, RMS, break index before, although they are also suggested as useful features in communicating importance, their effects were not verified by our data. We believe that our work can assist building a comprehensive prosody model for Concept-to-Speech generation and at the same time, help identify abnormality in utterances for speech analysis and understanding.

12. References

[1] Pan, S. and McKeown, K., "Learning intonation rules for concept to speech generation", Proc. COLING-ACL, Montreal, Canada, 1998.

[2] Pan, S. and McKeown, K., "Word informativeness and automatic pitch accent modeling", Proc. EMNLP, College Park, Maryland, 1999.

[3] Pan, S. and Hirschberg, J., "Modeling local context for pitch accent prediction", Proc. ACL, Hong Kong, 2000.

[4] Dalal, M. and Feiner, S. and McKeown, K. and Pan, S. and Zhou, M. and Hoellerer, T. and Shaw, J. and Feng, Y. and Fromer, J., "Negotiation for automated generation of temporal multimedia presentations", Proc. ACM MM, 1996.

[5] Silverman, K. and Beckman, M. and Pitrelli, J. and Ostendorf, M. and Wightman, C. and Price, P. and Pierrehumbert, J. and Hirschberg, J., "ToBI: a standard for labeling English prosody", Proc. ICSLP, 1992.

[6] Halliday, M., An introduction to functional grammar, Edward Arnold, London, 1985.

[8] Chambers, J. and Hastie, T. "Statistical models in S", Wadsworth & Brooks, Pacific Grove, California, 1992.

[9] Cahn, J., "Generation of affect in synthesized speech", J. AVIOS Soc., pp 1-19, 1990.

[10] Williams, W. and Stevens, K. N. , "Emotions and speech", J. Acoust. Soc. Amer., Vol. 52, n 4, pp 1238-1250, 1972.

[11] Prevost, S., "A semantics of contrast and information structure for specifying intonation in spoken language generation", PhD thesis, Univ. Pennsylvania, 1995.