AUTOMATIC DETECTION OF CONTRASTIVE ELEMENTS IN SPONTANEOUS SPEECH

*Ani Nenkova**

University of Pennsylvania
nenkova@seas.upenn.edu

Dan Jurafsky

Stanford University
jurafsky@stanford.edu

ABSTRACT

In natural speech people use different levels of prominence to signal which parts of an utterance are especially important. Contrastive elements are often produced with stronger than usual prominence and their presence modifies the meaning of the utterance in subtle but important ways. We use a richly annotated corpus of conversational speech to study the acoustic characteristics of contrastive elements and the differences between them and words at other levels of prominence. We report our results for automatic detection of contrastive elements based on acoustic and textual features, finding that a baseline predicting nouns and adjectives as contrastive performs on par with the best combination of features. We achieve a much better performance in a modified task of detecting contrastive elements among words that are predicted to bear pitch accent.

Index Terms— focus detection, contrastive elements, discourse understanding

1. INTRODUCTION

In natural speech people use a variety of prosodic means to convey to their interlocutor which elements of the utterance are especially important. Often the production of stronger than usual prominence is realized over appropriate words or phrases, making speech more expressive and signaling the *focus* [1, 2] of the utterance, where one *contrastive element* is chosen among a limited set of alternatives.

The most clear examples of focus contrastive elements are question-answer pairs in which the contrastive elements pick out an answer among a set of feasible other alternatives.

Q: What did you have for dinner?

A: SALMON, and a CHOCOLATE MOUSSE for desert.

Contrastive elements often occur outside of question-answer pairs as well, when the context of the utterance contains an explicit reference to a contrastive alternative as in the following examples.

1. It is not in SOUTH Asia, it's in EAST Asia.
2. I really LIKED the guy, but John SUSPECTED him of fraud.
3. Be careful with this plate, it is EXTREMELY hot.

The detection of contrastive elements constitutes an important subtask for automatic speech understanding and dialogue systems development, since it has to be taken into account when modeling the speakers attentional state and intentions.

The ability to produce contrastive elements is also important for text-to-speech. Perception experiments show that modeling regular prominence (pitch accent) and stronger prominence (emphatic accent) improves the quality of unit selection speech synthesis [3]. An automatic classifier for contrastive accent can be used to label the voice database for correctly synthesizing such accents. [1]

While contrastive speech is thus important for both speech recognition and synthesis, few studies have examined the characteristics of expressive contrastive accents in natural speech.

In the reminder of the paper we overview related work (Section 2) and describe our corpus of spontaneous dialogues labeled for contrastive elements (Section 3). Then we perform an analysis of the acoustic properties of contrastive elements and regular pitch accented words to verify the existence of salient differences between them in Section 4. In Section 5 we present our contrastive elements detector, and discuss our findings in Section 6.

2. RELATED WORK

Much of the previous work on detection of contrastive elements has been motivated by the need to improve naturalness and expressivity of the output of speech synthesis [5, 6, 7, 8]. These studies concentrate on the analysis of clear speech recorded in a studio by professional speakers. In [9] for instance, the same passages were recorded both in a neutral and contrastive context, e.g. "*We painted the house white.*" vs. "*We painted the barn red, but we painted the house white.*" In terms of TOBI labeling, contrastively emphasized words

*The author performed part of the work while a postdoctoral fellow at Stanford University.

[1] Previous research shows that the obvious alternative methods are both problematic: instructing the voice talent to read certain words with greater prominence leads to inconsistent strength of emphasis [4], and carefully constructing contrastive frames ("It wasn't X who did it, it was Y.") results in better quality contrasts but requires significant work in script construction.

were found to consistently have intermediate prosodic phrase boundaries on each side of the word. The words were marked with a high pitch accent H* and a low phrase accents L-. In a more recent study [10], an emphasis detector based on acoustic features was trained on the specially recorded emphatic corpus and used to label the main part of the synthesis database. For these recordings done in a controlled environment by professional speakers, the acoustic features lead to very good detector performance with f-measure of 0.8.

Numerous previous studies have also discussed the importance of focus detection for speech understanding. Some detection systems, for example, have been built for specific applications, such as child computer-based tutoring, in which the detection of the novel part of an utterance and of syntactically parallel contrastive elements was necessary for dialog understanding [11]. In other studies, fundamental frequency, phrase boundaries and sentence mode have been shown to be helpful for focus detection [12], as well as overall intensity and spectral tilt (for emphasis detector in Swedish) [13].

In a scenario closest to ours [14], Switchboard annotations were used to study within a solid theoretic framework how prominence and information structure align and to predict contrastive elements using features such as information status, syntactic category and manually labeled three way-prominence level (non-accented, non-nuclear pitch accent and nuclear pitch accent).

3. DATA AND FEATURES

For our study we used 12 Switchboard conversation that have been annotated for contrastive elements following the labeling framework outlined in [15]. This annotation scheme is based both on perceptual cues with annotators listening to the audio and on semantic theories of focus where direct contrast due to syntactic parallelism has been extended to incorporate a larger class of contrastive elements that pick out one of a set of possible alternatives. Elements not falling into any of the categories of contrastive elements are marked as *background*, or non-contrastive. Different subclasses of contrastive elements include answer (the phrase is an answer to a question), subset (entities that have a common supertype), contrastive (directly compared with an alternative in the utterance context), adverbial (word made contrastive by the use of a focus-inducing word such as "just" or "also"). In our study the different subclasses were not distinguished and were grouped together to form the class of contrastive elements.

Some parts of the conversations containing disfluent or highly ungrammatical utterances were not annotated and are excluded from our analysis. The final corpus contained 7,785 annotated words in total, 2,150 of which were marked as contrastive elements.

In addition the corpus has been manually annotated on the word level for the presence or absence of pitch accent.

Below are some examples from the corpus from a conversation about options for child care. Words in capital letters were produced as prominent by the speaker, and marked as bearing a pitch accent.

1. /my EXPERIENCE/$_{contrastive}$ is JUST with what WE$_{adverbial}$ did and so they DIDN'T really go through the /CHILD care ROUTE/$_{contrastive}$.
2. i have a /philosophical PROBLEM/$_{other}$ with THAT.
3. ... and DROP a /TWO year OLD/$_{subset}$ OFF in a HOME$_{contrastive}$ where you KNEW there were going to be /FOUR other KIDS/$_{subset}$.
4. (How much does a nanny cost?)
   i THINK it's about /SIXTY DOLLARS a WEEK for TWO children/$_{answer}$.
5. you$_{contrastive}$ TAKE this subject much more PERSONALLY$_{other}$ than I$_{contrastive}$ do.

The features we considered for the detection of contrastive elements included both acoustic and non-acoustic features. Fundamental frequency (f0) and energy features were extracted automatically for each word using Praat, and normalized by speaker.

F0 Minimum (pmin), maximum (pmax), range (pmax - pmin), average (pavr).

Energy Minimum (emin), maximum (emax), range (emax - emin), average (eavr) .

Duration Word duration extracted from the Mississippi State University Switchboard transcripts, not normalized.

Pause Length of pause after the word, based on the start and end time of words in the transcripts; not normalized.

Part-of-speech Six broad part of speech classes were considered: adjectives, adverbs, function words (prepositions and determiners), nouns, pronouns, verbs. Gold standard manual annotations were used.

Accent ratio This is a lexicalized feature that proved to be useful for pitch accent prediction [16, 17]. It takes values between 0 and 1 and is based on an accent ratio dictionary containing words that appeared in a larger corpus as either accented or non-accented significantly more often than chance. The value of the accent ratio feature is the probability of the word being accented if the word is in this pre-built dictionary and 0.5 otherwise.

Before turning to the task of detection of contrastive elements and non-contrastive elements, we first present a descriptive analysis and comparison between the contrastive, pitch accented and non-prominent words.

	contrastive	non contrastive
accented	1778	2320
non accented	372	3315

Table 1. Corpus distribution across accented and contrastive categories. The two are highly correlated, with contrastive elements predominantly bearing accent.

4. CONTRASTIVE ELEMENTS AND PITCH ACCENT

As a first step in our stud we first need to verify that contrastive elements in our corpus are acoustically different from regular pitch accent prominence. Since the corpus annotation guidelines combine in the definition of contrastive elements both semantic considerations and perceptual evidence, it is also important to confirm that there are salient difference between the classes of contrastive and pitch accented words.

4.1. Are most contrastive elements accented?

Under our working hypothesis that contrastive elements are more emphatic than other elements of the sentences, it is desirable that most contrastive elements are also accented. An additional requirement is that the contrastive element class is not equivalent to the class of pitch accented words, since the latter problem has already been extensively studied with very good results [18, 19, 20, 21, 22, 17]. The Switchboard data and annotations support both requirements.

Table 1 shows the distribution of words between the accented (bearing pitch accent) and contrastive categories. As expected, pitch accent and contrastive status are highly correlated and, specifically, there is a highly significant tendency for contrastive words to be accented—83%. Most of the remaining 17% of contrastive words that are not accented are part of the longer noun phrase that carries the contrast, like the unaccented "care" or "philosophical" in "CHILD care ROUTE" and "philosophical PROBLEM" in the corpus examples above.

At the same time, pitch accent is not that predictive of contrast status, with only 43% of all the accented tokens also being contrastive. This distribution indicates that contrastive elements form a special class of pitch accented items and that the two classes are not essentially the same. The task of detection of contrastive elements in conversation is clearly well-specified and different from the pitch accent prediction task.

4.2. Acoustic differences between contrastive, accented and non-prominent elements

As attested by Table 1, contrastive words do not coincide with the class of pitch accented words, even though contrastive words do tend to be predominantly accented. We now turn to examine the specific acoustic differences between contrastive words and words that bear pitch accent.

Table 2 shows the average values for the acoustic measurements related to pitch, intensity, duration and pause length. In addition, the last two columns in the table show the values for acoustic features for contrastive words that bear pitch accent versus words that bear a pitch accent but are not contrastive. This difference corresponds to the difference between regular pitch accent and the potentially more emphatic contrastive accent. As the table shows, there are salient difference between the two, and some of the differences in acoustic features are significant.

Table 3 gives the p-values (from a two-sided t-test) for difference in three comparisons: *(i)* no accent vs. pitch accent; *(ii)* pitch accent vs. contrast; *(iii)* contrast vs. no contrast; *(iv)* accent+contrast vs accent-contrast

As expected, the acoustic differences in comparison *(i)* between words bearing pitch accent and those that don't are all highly significant (first column in Table 3). Similarly, in comparison *(iii)*, contrastive and non-contrastive elements acoustically behave quite differently.

In comparison *(ii)* between pitch accent and contrast, the most salient significant difference is that of duration, with contrastive elements on average having longer duration than words bearing plain pitch accent. f0 minimum is also significantly different between contrastive and accented items, with, interestingly, average f0 higher for accented, not for contrastive, words.

Finally, we turn to the comparison between items that are both accented and contrastive and those that are accented but not contrastive (*(iv)*). It shows again that the contrastive distinction bears salient information beyond plain accenting. All three measures for f0 and energy—minimum, maximum, and range—are highly significantly different. In the conversational setting, speakers not only make contrastive elements prominent, but also use different acoustic realizations compared to those use to mark importance using pitch accent.

5. DETECTING CONTRASTIVE ELEMENTS

For our contrastive element detector we used the multinomial logistic regression model with a ridge estimator based on [23] in the WEKA toolkit [24]. The categorical part of speech feature was converted to six binary features, one for each broad part of speech class.

Table 4 shows the performance of the detector from 10-fold cross-validation using different features. The majority class (non contrastive) baseline gives 72.38% accuracy. Part of speech and accent ratio are the only features that used in isolation lead to improved accuracy over the baseline. The accent ratio feature and all acoustic features in combination have really low recall, leading to poor overall accuracy.

Surprisingly, using the six part of speech features leads to very good accuracy of 76.42%, and balanced and reasonable precision and recall. The detector based solely on part of speech features predicts that all nouns and adjectives are

	no accent	pitch accent	no contrast	contrast	accent-contrast	accent+contrast
pmin	-0.1073	-0.1158	-0.1055	-0.1282	-0.1057	-0.1290
pmax	0.0987	0.1612	0.1167	0.1709	0.1448	0.1827
prange	0.2061	0.2771	0.2222	0.2991	0.2505	0.3118
pavr	-0.0063	0.01512	0.0035	0.0085	0.0166	0.0131
emin	-0.0536	-0.1262	-0.0736	-0.1395	-0.1093	-0.1483
emax	0.4186	0.4880	0.4410	0.4922	0.4777	0.5015
erange	0.4722	0.6143	0.5147	0.6318	0.5871	0.6498
eavr	0.2348	0.2580	0.2447	0.2532	0.2584	0.2575
duration	0.1911	0.3495	0.2312	0.3879	0.3016	0.4119
pause	0.0306	0.0914	0.0502	0.0950	0.0822	0.1036

Table 2. Acoustic characteristics of contrastive, non-contrastive, accented and non-accented elements. Differences for all acoustic measures are significant between accented and non-accented elements, while the significant differences between contrastive and accented are only for minimum and range for f0 and energy.

	no accent vs. pitch accent	pitch accent vs contrast	no contrast vs. contrast	accent+contrast vs. accent-contrast
pmin	0.009746	0.001872	1.649e-09	7.375e-07
pmax	<2.2e-16	NS	5.148e-10	0.0002094
prange	<2.2e-16	0.02409	<2.2e-16	1.341e-08
pavr	1.060e-15	0.02075*	NS	NS
emin	<2.2e-16	NS	<2.2e-16	2.053e-05
emax	< 2.2e-16	NS	<2.2e-16	3.373e-06
erange	<2.2e-16	0.02687	<2.2e-16	8.612e-11
eavr	5.571e-10	NS	0.0347	NS
duration	<2.2e-16	<2.2e-16	<2.2e-16	< 2.2e-16
pause	<2.2e-16	NS	NS	NS

Table 3. P-values for the differences between classes according to the acoustic measures. NS = not significant, larger than 0.05.

	accuracy	precision	recall	f-measure
majority	72.38%	0	0	0
all features	76.88%	0.605	0.469	0.528
POS	76.42%	0.567	0.622	0.593
all acoustic	72.60%	0.529	0.073	0.128
accent ratio	72.56%	0.510	0.173	0.258

Table 4. Classifier performance; distinguishing contrastive and contrastive elements, regardless of accent.

	contrastive	non contrastive
ADJ	342	212
ADV	187	804
FUN	126	762
NN	995	811
PRO	122	1323
VB	378	1720

Table 5. Distribution of contrastive elements across broad part of speech.

contrastive, while the remaining broad part of speech classes are non contrastive. Indeed, the main function of these part of speech categories is to pick out one among possible alternatives for class and attributes of entities. Still, we do know that not all adjectives and nouns are contrastive, and that every other part of speech can carry contrastive meaning when given an appropriate context. Given that part of speech turned out to be such a strong baseline, one would be interested to see the distribution of contrastive elements across part of speech classes and this information is given in Table 5. About half of all nouns and verbs are contrastive, while for the other parts of speech contrastive elements occur rarely.

In general it was to be expected that results on this corpus of spontaneous conversational speech results will be lower than those for professionally read speech (cf. [10]). But such detection results are generally disappointing, especially given that acoustic features do not seem to be helpful for detecting contrastive elements. Our descriptive analysis showed that there are highly significant differences between contrastive and non contrastive elements. One reason why these differences are not as helpful in the general contrast detection task is that there is a imbalance between the two classes, with three-fourths of all words being non contrastive; the difference in means, while highly significant, is not sufficient to keep the number of false contrastive positives smaller than falsely predicted non contrastive elements in order to optimize a classifier. In addition, the classifier is presumably confused by the non-accented contrastive words (like *care* and *philosophical* discussed in Section 4.1, which will have a very

	accuracy	precision	recall	f-measure
majority	58.08%	0	0	0
all features	68.38%	0.643	0.552	0.594
all acoustic	66.03%	0.642	0.428	0.513
POS	65.59%	0.576	0.683	0.625
duration	65.43%	0.635	0.413	0.500
erange	58.77%	0.524	0.180	0.268
accent ratio	57.60%	0.477	0.120	0.192

Table 6. Prediction results; distinguishing accented contrastive elements from accented non-contrastive.

different distribution of acoustic features than accented contrastive words, rendering the acoustic features less reliable.

5.1. Can we distinguish accented contrastive from accented non-contrastive?

The previous section suggested that non-accented contrastive items (like *care*), while relatively rare, may have kept acoustic features from helping our first contrast detector. Recall also our earlier analysis that showed that pitch-accented elements that are contrastive have more extreme acoustic properties than pitch-accented words that are not contrastive. These two results suggest that we attempt to *detect contrast only among the pitch-accented words*. This new task is also natural given our original goal for TTS of distinguishing between regular pitch accents and emphatic accents for labeling the speech synthesis database, and is even more compelling given that in our corpus only 154 words among those predicted as not bearing pitch accent are contrastive.

A classifier for contrastive elements was trained and tested only on the portion of the data automatically predicted by a pitch accent classifier based on the accent ratio feature to be accented. This classifier predicts that all words with accent ratio smaller than 0.38 are not accented, and all other words are accented. The contrast detection results in this setting are better than those for a general contrastive detector. Notably, the baseline in the modified task is lower (58.08% accuracy), and the relative difference between the detectors and the baseline is considerably larger. Moreover, in this reduced classifier, some contrastive and some non-contrastive elements were predicted for *each* part of speech, including contrastive elements for pronouns and function words.

The majority class baseline for the modified task is 58.08% accuracy. Part of speech is still the best single feature for the detection of contrastive accent, with 65.59% accuracy. But now the combination of all acoustic features has better accuracy and precision than part of speech, showing that the statistically significant differences in means discussed in Table 3 can be reliably exploited for building a detector of accented contrastive elements. The overall best detector is the one in which both acoustic and non-acoustic features are combined to reach accuracy of 68.38%.

In our descriptive analysis of contrast and pitch accent we saw that these two prominence dimensions are highly correlated but not overlapping. Now, we see that the contrast detection task in isolation is difficult and the best performance is that of a reasonable baseline based on noun or adjective versus other part of speech distinction. When pitch accent is incorporated in the contrastive element detection class, performing the detection only over the words that have already been predicted to be accented, detection performance increases: the difference between the baselines and the best classifier grows, as well as the power of a detector based solely on acoustic features.

This difference in performance requires further investigation in the future. One very likely explanation comes from the pronounced difference between words with no pitch accent and those with pitch accent. These two classes are acoustically very different, and the complete contrastive class contains some elements from both, possibly introducing confounds between the classes. Alternatively, the pitch accent predictor based on accent ratio eliminates some of the hard cases of contrast by predicting them as not bearing pitch accent.

6. DISCUSSION

In this paper we have presented a study of the acoustic correlates of contrastive accent in conversational speech. In a descriptive comparison between contrastive elements and pitch accented words we found significant acoustic differences in duration, f0 minimum, as well as in f0 and energy range, and f0 average. There are more significant differences in acoustic features between items that bear both pitch accent and contrast compared to those with pitch accent but no contrast.

Results from a contrast vs. no contrast detector show that for spontaneous conversational speech acoustic features are not sufficient to make this distinction reliably. A detector predicting that nouns and adjectives are contrastive and all other words non-contrastive achieves the best balance in terms of precision and recall. Such results give a competitive baseline for future studies on contrast detection. They also suggest an approach for synthesis of contrastive elements: possibly pronouns and prepositions synthesized from units coming from nouns and adjectives will indeed sound more emphatic. This hypothesis should be tested in perception experiments. We suggest that the fact that acoustic features don't help in this task may be due to the confound introduced by unaccented words inside contrastive noun phrases.

Finally, much better results were achieved in a task of identifying which elements are contrastive among those that are already predicted to bear pitch accent. In this task, acoustic features proved to be reliable indicators and led to good detector performance, as well as overall greater improvement over the baseline.

7. ACKNOWLEDGEMENTS

We are immensely grateful to Sasha Calhoun for providing us with the NXT Switchboard corpus and for insightful discussions. We are also thankful to Vivek Kumar for the extraction of acoustic features used in this study. This work was supported by the Stanford LINK and the ONR (MURI award N000140510388).

8. REFERENCES

[1] Mats Rooth, "A theory of focus interpretation," *Natural Language Semantics*, vol. 1, no. 1, pp. 75–116, 1992.

[2] E. Vallduví and M. Vilkuna, "On rheme and kontrast," *Syntax and Semantics*, vol. 29, pp. 79–108, 1998.

[3] Volker Strom, Ani Nenkova, Robert Clark, Yolanda Vazquez-Alvarez, Jason Brenier, Simon King, and Dan Jurafsky, "Modelling prominence and emphasis improves unit-selection synthesis," in *Proceedings of Interspeech*, Antwerp, Belgium, 2007.

[4] Volker Strom, Robert Clark, and Simon King, "Expressive prosody for unit-selection speech synthesis," in *Proceedings of Interspeech*, Pittsburgh, USA, 2006.

[5] S. Prevost, *A Semantics of Contrast and Information Structure for Specifying Intonation in Spoken Language Generation*, Ph.D. thesis, University of Pennsylvania, 1995.

[6] Mariët Theune, "Contrastive accent in data-to-speech system," in *ACL-EACL*, 1997, pp. 519–521.

[7] M. Wolters and P. Wagner, "Focus perception and prominence," in *Proceedings of the 4th Conference on Natural Language Processing KONVENS-98; Computer Studies in Language and Speech Vol. 1*, 1998, pp. 227–236.

[8] Bettina Braun and D. Robert Ladd, "Prosodic correlates of contrastive and non-contrastive themes in german," in *Proceedings of the 8th European Conference on Speech Communication and Technology*, Geneva, 2003, pp. 789–792.

[9] John F. Pitrelli and Ellen M. Eide, "Expressive speech synthesis using american english tobi: questions and contrastive emphasis," in *IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU'03*, 2003.

[10] Raul Fernandez and Bhuvana Ramabhadran, "Automatic exploration of corpus-specific properties for expressive text-to-speech: A case study in emphasis," in *6th ISCA Workshop on Speech Synthesis*, 2007.

[11] Tong Zhang, Mark Hasegawa-Johnson, and Stephen E. Levinson, "Extraction of pragmatic and semantic salience from spontaneous spoken english," *Speech Communication*, vol. 48, pp. 437–462, 2006.

[12] Anja Elsner, "Focus detection with additional information of phrase boundaries and sentence mode," in *Proceedings of EUROSPEECH-1997*, 1997, pp. 227–230.

[13] M. Heldner, E. Strangert, and T. Deschamps, "A focus detector using overall intensity and high frequency emphasis," in *Proceedings of ICPhS-99*, 1999.

[14] Sasha Calhoun, "Predicting focus through prominence structure," in *Proceedings of Interspeech'07*, 2007.

[15] S. Calhoun, M. Nissim, M. Steedman, and J.M. Brenier, "A framework for annotating information structure in discourse," *Pie in the Sky: Proceedings of the workshop, ACL*, pp. 45–52, 2005.

[16] J. Brenier, A. Nenkova, A. Kothari, L. Whitton, D. Beaver, and D. Jurafsky, "The (non)utility of linguistic features for predicting prominence in spontaneous speech," in *IEEE/ACL 2006 Workshop on Spoken Language Technology*, 2006.

[17] Ani Nenkova, Jason Brenier, Anubha Kothari, Sasha Calhoun, Laura Whitton, David Beaver, and Dan Jurafsky, "To memorize or to predict: Prominence labeling in conversational speech," in *Proceedings of NAACL-HLT*, Rochester, NY, 2007.

[18] J. Hirschberg, "Pitch accent in context: Predicting intonational prominence from text," *Artificial Intelligence*, vol. 63, 1993.

[19] K. Ross and M. Ostendorf, "A dynamical system model for recognizing intonation patterns," in *Proceedings of Eurospeech*, Madrid, Spain, 1995, pp. 993–996.

[20] Shimei Pan and Julia Hirschberg, "Modeling local context for pitch accent prediction," in *Proceedings of ACL'00*, Hong Kong, 2000, pp. 233–240.

[21] X. Sun, "Pitch accent prediction using ensemble machine learning," in *Proceedings of ICSLP*, 2002.

[22] Michelle Gregory and Yasemin Altun, "Using conditional random fields to predict pitch accents in conversational speech," in *Proceedings of ACL'04*, 2004.

[23] S. Le Cessie and J. C. Van Houwelingen, "Ridge estimators in logistic regression," *Applied Statistics*, vol. 41, pp. 191–201, 1992.

[24] G. Holmes, A. Donkin, and I. Witten, "Weka: A machine learning workbench," in *Second Australian and New Zealand Conference on Intelligent Information Systems*, 1994.