Automatic Detection of Contrast for Speech Understanding

Tong Zhang, Mark Hasegawa-Johnson, and Stephen E. Levinson

Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign
{tzhang1, hasegawa, sel}@ifp.uiuc.edu

Abstract

Contrast is a very popular phenomenon in spoken language, and carries very important information to help understanding contents and structures of spoken language. In this paper, we propose an idea of automatic contrast detection as an effort for better speech understanding. We study the automatic tagging of three specific types of contrast: symmetric contrast, contrastive focus, and contrastive topic. We label the three types of contrasted words as contrast (C), and other words as noncontrast (¬C). The classification of contrast events is based on prosodic, spectral, and part-of-speech (POS) information sources. The integration of different knowledge sources is realized by a time-delay recursive neural network (TDRNN). The approach we proposed was testified on 235 spontaneous utterances consisting of 3500 words (samples). The contrast detection was speaker independent. The tests yielded an average of 87.9% classification rate.

1. Introduction

Contrast is a complicated concept which has been defined from different perspectives by Linguistic and Psychological researchers. For example, one definition originates from the logical notion of contrariety [1]. Contrariety defines two propositions to be contrastive if it is impossible for them to be true simultaneously. For example, in the sentence "Bach was an organ mechanic; Mozart knew little about organs", the two propositions are not contrastive, whereas become contrastive when "Bach" is replaced by "Mozart" at the beginning of the second sentence. Another definition is based on syntactic parallelism [2]. For example in the sentence "an American farmer talked to a Canadian farmer", words "American" and "Canadian" are contrasted. In addition, there is also an opinion which defines contrast as novelty in the sense that novelty usually conveys a contrast between a fact and the potential alternatives [3].

Despite of disputes about the academic definition and other issues of contrast, we intend to apply some study results on contrast which have been achieved to pragmatically automatic speech understanding. In this paper, we investigate three sub-topics of contrast: symmetric contrast, contrastive focus and contrastive topic. The three sub-topics are relatively well formalized although disparity still exists on how to define them accurately. Here we give our definition by which we collect data for our statistical study. Our definition refers to publications of linguists on this issue.

- Symmetric contrast consists of a set of two or more distinct words which are parallel in syntactic structure, and the emphasis on one word is motivated by the contrast with the other word(s) [3]. We take the following sentences for instance (contrasted words are denoted using **bold**),

  (1) *Show me flights arriving into **Baltimore** by 10pm from **Denver**, no, from **Chicago**.*
  (2) ***This** one will take longer than **that** one.*

- Contrastive focus marks something new and contrasted with presuppositions [4]. Focus is interpreted as representing the syntactic constituent which forms a novel assertion, whereas the rest of the sentence is presupposed by the listener. In the contrastive case, the novel assertion corrects an explicit or implicit assumption made by the listener. In the following two examples, "this" corrects "that", and "doubt" corrects "dad".

  (1) *A: Take that big gear, please.*
  *B: I thought you said **this** gear.*
  (2) *B: I'd say I doubt it would work.*
  A: With your dad, it would work.
  *B: I **doubt** it.*

- The subjects of two conjunct predictions constitute contrastive topic if the conjuncts contains opposite information to each other. The conjuncts may be connected by either *and* or *but*, indicating either a parallel or a contrastive discourse relation [3]. Unlike contrastive focus, contrastive topic is usually the presupposed part of the sentence. For example,

  (1) *A: Where are the red gear and the yellow gear?*
  *B: The **red** gear is on the bottom, and the **yellow** gear is on the top.*
  (2) *A: How are the gears spinning?*
  *B: The two **outside** ones spin in the same direction and the **middle** one spins in the opposite direction.*

Contrast is very important in the analysis of information structure, since contrasted words usually contain new and important information and are more easily to be recognized because of pronunciation emphasis. Contrast detection helps better interpretation and rich transcription of speech data. To date the automatic transcription of spontaneous speech has involved dysfluency [5][6], intonational phrases and discourse markers [7], punctuation [8], turn boundaries [9] [10], and dialogue act [9][11]. However, to our knowledge contrast annotation has not been investigated yet. In this paper, we study how to automatically annotate contrasted words in speech.

2. Data corpora

2.1. ATIS0

The ATIS0 is a data corpus distributed by the Linguistic Data Consortium (LDC). It is collected to develop a conversationally proficient airline information assistant, which helps a user to make a travel schedule. It consists of 912 utterances elicited by 36 speakers and collected by Wizard-of-Oz (WoZ) dialogues.

2.2. ITS

The ITS on which we are working is an intelligent tutoring system used to help children learn some basic concepts of Mathematics and Physics. Children can acquire knowledge through manipulating concrete objects (Legos) rather than solely handling abstract symbols [12]. In our WoZ experiments, the children are given gears of different sizes. The teeth on each gear are painted with different color pairs: red and blue, red and green, or blue and green. The tutor helps children by asking children questions, guiding them to use Legos to find solutions of the questions, and answering questions which they propose. Meanwhile, the tutor provides emotional support and consolation, and carefully adjusts his tutorial strategy according to emotion and learning progress of the children. For example, one question is about the ratio of the teeth number and spinning cycles:

> *Line up a 24-tooth gear and a 40-tooth gear. If the 24-tooth gear spins 5 times, then how many times must the 40-tooth gear spin for them to line up again? Why?*

Children are expected to line up a 24-tooth gear and a 40-tooth gear along a beam and right next to each other, and then rotate the gears, counting and comparing the spinning cycles. Children should observe that gears with more teeth spin more slowly. Some children further discover that the product of teeth number and spinning cycles is the same for the two gears.

To date 714 students' utterances have been collected, containing approximately 50mins of relatively clean speech. On average each utterance had 4.2s speech and 8.1 words. Of these utterances, we deleted those utterances which were meaningless such as "Uhm … like…", and those were not containing a complete semantic unit such as "When I do like…", and use the other utterances for experiments.

2.3. Annotation

Two students annotated independently of each other on those utterances which they thought to contain contrast examples. The annotation was performed based on speech perception, transcription, and dialogue context. Then they talked with each other, and finally reached an agreement on 267 utterances about contrast. We found that in ATIS0 many sentences containing symmetric contrast had a same syntactic structure "from … to …" around the contrasted words, such as "from Philadelphia to Denver." To avoid the data monotone in syntactic structures, we selected only a few instances with such structure. Finally we obtained 235 utterances containing contrast examples.

3. Proposed method

3.1. Information sources

3.3.1 Prosody

Prosody captures paralinguistic information by looking into the aspects of speech signals other than actual words spoken. An important attribute of contrast is pitch accent. Although contrast remains a problematic notion, it is a unanimous opinion that contrast is accented in speech [1]. Pitch accent is usually detected using pitch and energy. Pitch and energy are extracted using the "formant" program in Entropic XWAVES with a probability of voicing (PV) that serves as a confidence measure. Then pitch is normalized to compensate for the difference in pitch range across speakers by:

$$pitch_f = \frac{pitch_f - \min_f(pitch_f)}{\max_f(pitch_f) - \min_f(pitch_f)} \quad (1)$$

where pitch_f is the pitch value of the f^{th} frame, $\min_f(pitch_f)$ is the minimum non-zero pitch value of the entire utterance, and $\max_f(pitch_f)$ is the maximum pitch value of the entire utterance. Energy is normalized by the peak value in order to compensate for the differences in the sound volume across speakers.

The frame-level pitch and energy are averaged in a special way to obtain the syllable-level feature vector. The averaging scheme is given by:

$$D_m = \frac{1}{N_{\psi_m}} \sum_{f \in \psi_m} F_m[f] \quad (2)$$

where D_m is the feature vector for syllable S_m, $F_m[f]$ is the feature vector of frame f in S_m, ψ_m is a subset of frames in S_m, N_{ψ_m} is the total number of frames in ψ_m, and

$$\psi_m = \{ f \mid |F_m[f]| \geq \frac{1}{2} T_m \}, \quad (3)$$

$$T_m = \max_{F_m[f] \in S_m} \{F_m[f]\}. \quad (4)$$

Then we take the maximal pitch and energy of syllables in a word as the word-level pitch and energy features.

In addition, studies have shown that word duration is affected by its occurrence frequency in discourse, and its predictability from its following word [13]. Novel words tend to have longer duration than presupposed words, and content words tend to have longer duration than function words. Contrasted words are usually content words expressing novel or important information. Therefore, duration is a useful attribute for contrast detection. We use forced alignment to determine word boundaries, and phoneme boundaries whereby to derive syllable boundaries. The maximal syllable duration in a word is used as a representation of the word duration.

3.3.2 Spectral balance

We use spectral balance to capture the spectral characteristics of contrastive stress. Spectral balance is defined as the intensity increase at higher frequencies (≥ 500 Hz) of vocal speech. Perceptual experiments showed that spectral balance

was a reliable indicator of stress [14]. If a speaker produced stressed syllables, then the intensity of signals at higher frequencies increased more than the intensity of signals at lower frequencies. The intensity level manipulation of signals at higher frequencies provided stronger stress than the manipulation of the entire frequency band. In spoken English, contrast usually occurs on the lexically stressed syllables which are produced with greater vocal effort. Therefore, spectral balance can be used as a spectral attribute for contrast detection. In order to apply spectral balance to our contrast detection, we extract features called spectral balance-based cepstral coefficients (SBCC). We first use Daubechies-4 wavelet filter to decompose time-domain speech signals into *N* bands. We next compute the signal intensity in each band. Discrete cosine transformation (DCT) is then applied to the intensity of bands to derive SBCC. The detailed derivation of SBCC is described in [16]. Similar to pitch, we derive the syllable-level SBCC using the method addressed in Section 3.3.1. Then we take the maximal SBCC of syllables in a word as the word-level SBCC. Here the dimension of SBCC is 13.

3.3.3 Part-of-speech tagging

POS tagging has been widely used for rich annotation such as repairs and discourse markers [7]. Since contrast usually falls on content words rather than function words, the POS tag of a word also provides a rough estimate of the probability that the word carries contrast information. POS tagging is automatically performed by Roth's tagger [15].

TDRNN requires the feature variables to be continuous or discrete, so POS must be converted from a character variable to an indicator variable. Here we have 31 POS tags, and we try two transformation strategies for the m^{th} POS tag ($1 \le m \le 31$):
(1) 5 binary variables whose decimal value equals *m*;
(2) 31 binary variables in which only the m^{th} variable is 1.

The features used in this study are summarized in Table 1.

Table 1: List of features defined on a word

Feature	Description
Abs_dur	Absolute duration of the word.
max_syl_dur	Maximal duration of the syllables in the word.
max_syl_f0	Maximal pitch of the syllables in the word.
max_syl_egy	Maximal energy of the syllables in the word.
max_sbcc	Maximal sbcc of the syllables in the word.
POS	The part-of-speech tag of the word.

3.2. Information fusion

A TDRNN is trained to classify contrast events in a similar way as the pitch accent detection described in [16]. TDRNN is a 4-layer back-propagation network with two recursive context layers, which feed back delayed values from the output layer (the 4^{th} layer) and its previous hidden layer (the 3^{rd} layer), respectively. The recurrent circuits are used to capture the contextual information, because the contrast annotation of a word affects and also is affected by the annotation of its neighbors. For example, it is rare to have two contrast events be adjacent to each other. In addition, TDRNN uses delayed inputs to capture the dependence of human perception on the spectral change and dynamics of speech signals.

4. Results and Discussion

In total we have collected 235 utterances which consisted of approximately 20 minutes of speech and 3500 words (contrast / noncontrast samples). We used approximately 90% of the utterances (about 3000 Samples) for training, and used the other 10% for testing. Contrast detection was speaker independent. First, we investigated contrast detection using combined information sources. The first feature combination contained the 5-variable POS, and the second feature combination contained the 31-variable POS. An average classification rate of 87.9% was achieved for the first feature combination, and 85.4% for the second feature combination. Table 2 lists the test results in precision, recall and *F*-score. We use the *F*-score set (contrast *F*-score and noncontrast *F*-score) for comparison. The test results shows that the first feature combination outperforms the second feature combination. Second, we investigated contrast detection using individual information sources, and list the results in Table 3. The 31-variable POS shows superiority to any other feature in contrast detection. The unexpectedly low contribution of pitch is probably due to inaccuracies in feature derivation caused by pitch doubling and halving.

In addition, the comparison between Tables 2 and 3 shows that the 31-variable POS outperforms the first feature combination, while the first feature combination outperforms the second feature combination. Our explanation is that although the 31-variable POS is an efficient feature, when it is combined with other features, the large input (1+1+1+1+13+31=48) of the TDRNN makes the function to be approximated very complex. The back propagation algorithm is hard to converge, and thus solution found is worse.

Table 2: Precision *p*, recall *r*, and F-score *f* using the combined information sources; POS has two representations

		p	***r***	***f***
combined; 5-variable POS	Contrast	0.620	0.733	0.672
	Noncontrast	0.944	0.909	0.926
combined; 31-variable POS	Contrast	0.611	0.212	0.314
	Noncontrast	0.868	0.975	0.918

Table 3: Precision *p*, recall *r*, and F-score *f* using the individual information sources

		p	***r***	***f***
Pitch	Contrast	0.310	0.367	0.336
	Noncontrast	0.867	0.834	0.850
Energy	Contrast	0.427	0.533	0.474
	Noncontrast	0.900	0.855	0.877
Duration	Contrast	0.460	0.483	0.472
	Noncontrast	0.894	0.885	0.890
Spectra	Contrast	0.511	0.783	0.618
	Noncontrast	0.951	0.848	0.896
5-variable POS	Contrast	0.622	0.467	0.533
	Noncontrast	0.897	0.943	0.919
31-variable POS	Contrast	0.619	0.750	0.678
	Noncontrast	0.951	0.913	0.932

To data the subject of our study has been only those utterances which contain contrast examples. In the future, we shall expand our investigation subject to cover all utterances in the data corpora. We shall also use word semantic analysis as an information source for contrast detection.

5. Conclusions

We have described the automatic annotation of three types of contrast, consisting of symmetric contrast, contrastive focus, and contrastive topic. The contrast detection was based on a TDRNN which combined the prosodic, spectral, and POS information sources. We annotated 235 utterances containing contrast samples from two WoZ data corpora for experiments. The utterances under study consisted of 3500 contrast / noncontrast word samples. We used 90% of the samples for training, and used the other 10% for testing. The test yielded an average of 87.9% classification accuracy. In the future, we shall add word semantic analysis, and expand the investigation subject to cover all of the utterances in the data corpora.

6. Acknowledgements

This work is sponsored by NSF grant number 085980. Statements in this paper reflect the opinions and conclusions of the authors, and are not endorsed by the NSF.

7. References

[1] Bosch P., and van der Sandt, R. *Focus: Linguistic, Congnitive, and Computational Perspectives.* Cambridge University Press, Cambridge, UK, 1999.

[2] H. Prüst. *On Discourse Structuring, VP Anaphora and Gapping*. PhD Dissertation, University of Amsterdam, 1992.

[3] Umbach, C. "On the notion of contrast in information structure and discourse structure", *Journal of Semantics*. To appear.

[4] Zubizarreta, M. L. *Prosody, Focus, and Word Order*. The MIT Press, Cambridge, MA, 1998.

[5] Shriberg, E., Bates, R., and Stolcke, A. "A prosody-only decision-tree model for disfluency detection", *Proc. Eurospeech'97*, 5:2383-2386, 1997.

[6] Baron, D., Shriberg, E., and Stolcke, A. "Automatic punctuation and disfluency detection in multi-party meeting using prosodic and lexical cues", *Proc. ICSLP, 2002*.

[7] Heeman, P. A. and Allen, J. F. "Speech repairs, intonational phrases and discourse markers: modeling speakers' utterances in spoken dialogue", *Computational Linguistics*, 25(4):1-45, 1999.

[8] Kim J.-H. and Woodland, P. C. "A combined punctuation generation and speech recognition system and its performance enhancement using prosody", *Speech Communication*, 41:563-577, 2003.

[9] Nöth, E., Batliner, A., Kieβling, A., Kompe, R., and Niemann, H. "Verbmobil: the use of prosody in the linguistic components of a speech understanding system", *IEEE Trans. on Speech and Audio Porcessing*. 8(5):519-532, 2000.

[10] Stolcke, A. "Modeling linguistic segment and turn boundaries for N-best rescoring of spontaneous speech", *Eurospeech*, 1997.

[11] Jurafsky, D., Bates, R., Coccaro, N., Martin, R., Meteer, M., Ries, K., Shriberg, E., Stolcke, A., Taylor, P., and Ess-Dykema, C. V. "Switchboard discourse language modelling project final report", *Johns Hopkins LVCSR Workshop*, 1997.

[12] Wilensky, U. "Abstract meditaions on the concrete", (I. Harel & S. Papert, ed.) *Constructionism*, Ablex, Norwood, NJ, 1991.

[13] Bell, A., Gregory, M. L., Brenier, J. M., Jurafsky, D., Ikeno, A., and Girand, C. "Which predictability measures affect content word durations?" *Proc. ISCA Workshop on Pronunciation Modeling and Lexical Access*, 2002.

[14] Sluijter, A., van Heuven, V. J., and Pacilly, J. "Spectral balance as a cue in the perception of linguistic stress", *The Journal of the Acoustical Society of America*, 101(1):503-513, 1997.

[15] Roth, D. and Zelenko, D. "Part of speech tagging using a network of linear separators", *COLING-ACL*, 1998.

[16] Zhang, T., Kim, S.-S., Hasegawa-Johnson, M., and Cole, J. "Speaker-independent automatic labeling of pitch accent", *Intl. Conf. on Speech Prosody*, 2004.