AUTOMATIC DETECTION OF CONTRASTIVE WORD PAIRS USING TEXTUAL AND ACOUSTIC FEATURES

Xiao Zang [1,2], *Zhiyong Wu* [1,2,3], *Yishuang Ning* [1,2], *Helen Meng* [1,3], *Lianhong Cai* [1, 2]

[1] Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems
Graduate School at Shenzhen, Tsinghua University, Shenzhen 518055, China
[2] Tsinghua National Laboratory for Information Science and Technology (TNList)
Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
[3] Department of Systems Engineering and Engineering Management
The Chinese University of Hong Kong, Shatin, N.T., Hong Kong SAR, China
zangxiaocs@163.com, zywu@sz.tsinghua.edu.cn, ningyishuang@126.com, hmmeng@se.cuhk.edu.hk, clh-dcs@tsinghua.edu.cn

ABSTRACT

Labeling emphatic words from speech recordings plays an important role in building speech corpus for expressive speech synthesis. People generally pronounce some words stronger than usual, making the speech more expressive and signaling the focus of the sentence. Contrastive word pairs are often pronounced with stronger prominences and their presence modifies the meaning of the utterance in subtle but important ways. We used a subset of Switchboard corpus to study the acoustic characteristics of contrastive word pairs and the differences between contrastive and non-contrastive words. To address the problem of automatic detection of contrastive word pairs, support vector machines (SVMs) are used to automatically detect contrastive word pairs. We report the results for automatic detection of contrastive word pairs based on textual and acoustic features. By adding acoustic features, a much better performance is achieved.

Index Terms— Automatic detection, contrastive word pair, acoustic features, support vector machines (SVMs)

1. INTRODUCTION

Emphasis plays a very important role in expressive speech synthesis to highlight the focus of an utterance to draw the attention of the listener. People use different prosodic means to instruct listeners the focus of sentence in natural speech. They often pronounce some words stronger than usual, making the speech expressive and signaling the focus of the sentence [1]. Emphasis improves the overall perception of synthetic speech [2] as appropriate assignment of emphasis improves the expressivity and naturalness.

Emphatic speech synthesis usually depends on the availability of emphasis speech corpora with appropriate emphasis labeling information. The construction of speech corpus is of great importance, but the workload of building such emphasis speech corpus manually is extremely huge. Some automatic methods should be introduced.

In this study, we focus on the detection of contrastive word pairs. The definition of contrastive word pairs is: an information structure relation that links two semantically related words that explicitly contrast with each other. Some examples of contrastive word pairs from the Switchboard corpus [3] are shown as follows:

a) One was a ***skirt***, and one was a ***pant***.
b) We have to separate our ***papers*** and our ***glass***.
c) I think you could recover from a ***pistol***, but not from a ***gun***.

where “skirt” contrasts “pant”, “papers” contrasts “glass”, and “pistol” contrasts “gun”.

Automatic detection of contrastive word pairs also has a number of applications in human language technology systems, including generating improved prosody contours in expressive text-to-speech (TTS) synthesis, content spotting in spoken language summarization systems, identification of focal words in speech understanding systems, and improved facial animation generation for interactive tutors.

Regarding the researches on automatic detection of contrastive word pairs, [4] proposes a combined use of acoustic features (energy, duration, F0, etc.), part-of-speech (POS) and semantic dissimilarity measure to automatically identify symmetric contrast, which consists of a pair of words that are parallel or symmetric in linguistic structure

978-1-4799-2186-7/14/$31.00 ©2014 IEEE

but distinct or contrastive in meaning. In [5], acoustic and lexical features are used to detect different classes of focus.

The latest and most relevant to our work on automatic detection of contrastive word pairs are [6][7]. In [6] a rich set of features including lexical, deeper syntactic and semantic features are used to recognize contrast. Good performance is achieved by combining these textual features. In [7], by adding accent ratio and word identity to other textual features that used in [6], a better performance was achieved.

Relying only on textual features is obviously not enough to identify a contrastive relation between two words, acoustic features are necessary to be proposed especially in the scenario of automatic speech corpus annotation.

The rest of this paper is organized as follows. In the remainder of the paper we present our contrastive word pairs detector, describe the modifications compared to [7], and report the results in Section 2. In the final section, we draw conclusions from our experiments and describe future directions for our research.

2. CONTRASTIVE WORD PAIR DETECTION WITH SUPPORT VECTOR MACHINE

We used support vector machines (SVMs) [14] as the tagger for automatic detection of contrastive word pairs given the text prompts and corresponding speech recordings. Textual features including lexical, syntactic and semantic features are derived from the input text. In addition to the textual features, we also proposed acoustic features (F0 min, max, mean, energy max, duration) for the task.

2.1. Speech Data

Our experiment used a subset of the Switchboard corpus [3] that had been annotated with syntactic structure [8] and information structure [9]. We selected sentences containing just one contrastive word pairs. Since our tagger relies on textual features only and doesn't consider the discourse context outside the sentence, we removed all the contrast relations that are not identifiable by simply looking at text.

To simplify the description, we will refer to the two words of each contrastive word pair as W1 and W2, where W1 precedes W2 in sentence. For each sentence both positive and negative examples of contrast are extracted. All word pairs sharing the same broad POS are extracted and then assigned +1 (positive) if the word pair is linked with contrast or -1 (negative) otherwise. An example is shown in Fig. 1 for the sentence of "One was a shirt, and one was a pant", where "shirt" and "pant" forms contrastive word pair.

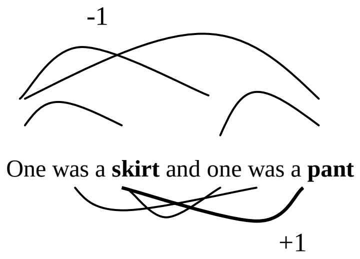

Fig. 1. Contrast example value generation from sentence: One/NN was/VBD a/DT skirt/NN and/CC one/NN was/VBD a/DT pant/NN, where "shirt" and "pant" are contrastive words. The contrastive word pair (skirt - pant) is given value +1 (i.e. positive example). All the other possible pairs of words sharing same broad POS are given value -1 (i.e. negative example).

2.2. Features

2.2.1. Common textual features

The features considered for detection of contrastive word pairs include all features as mentioned in [7]. These features were text-based and could be grouped into three categories: lexical features, syntactic features and semantic features.

Examples of **lexical features** are:

- Accent ratio that is the estimated probability of the word being accented in a training corpus.
- Word identity that refers to the English word itself.
- Single words or bigrams that activate contrast like "or", "rather than" in sentence.
- Textual similarity between two clauses containing W1 and W2.

Examples of **syntactic features** are:

- Dependency relation: If W1 and W2 have the same type of dependency relation (subject of, object of, etc.) with their heads (as in example b, both "you" and the first "I" have a "subject of" dependency with "take" and "do").
- If W1 is the only word having the same broad POS as W2 in sentence.

Examples of **semantic features** are:

- The semantic features consist of features indicating if W1 and W2 were linked by one of the following semantic relation: hypernyms, antonyms, entails, member-of, part-of, sisters.

2.2.2. New acoustic features

Previous research on the detection and perception of prosody has instructed that acoustic features based on f0, duration, and intensity were all indicators of prosodic prominence [13]. Besides, changes in pitch range have been found to signal a distinction between normal and emphatic accents [11].

We performed statistics of distribution of acoustic characteristics between contrastive words and non-contrastive words for sentences from the corpus, as shown in table I. As expected, all distribution for f0, energy and duration was apparent different. In comparison, contrastive words have higher f0, energy and longer duration. The most significant difference is duration, and contrastive word's duration is almost two times that of non-contrastive word in general.

Acoustic features including f0, energy and duration were considered for the detection of contrastive word pairs in our system, basing on above findings. Features were extracted for each word using Praat sound analysis package [12], and normalized. A variety of acoustic features for each word in our corpus were extracted. Five base acoustic features (f0 min, f0 max, f0 mean, energy max, and duration) were used in our experiments. Variants of each of these features were added to the feature set after normalization. We also included values for the raw and normalized features of the immediately neighboring words. Our final acoustic input vector contained 30 features: Five (5) basic features times three (3) words (current, previous, and succeeding), times two (2) normalizations (un-normalized, normalized by sentence).

Table I. Acoustic characteristics of contrastive and non-contrastive words. Differences of all characteristics measures are significant between contrastive and non-contrastive words, especially for duration.

	Non contrastive	Contrastive
F0 Min	0.1058	0.1286
F0 Max	0.1260	0.1701
F0 Mean	0.1103	0.1488
Energy Max	0.383034	0.431678
Duration	0.23828	0.442083

2.2.3. Summary of the features

All the features described below were used in our SVM tagger for automatic detection of contrastive word pairs.

- **Accent ratio**: This is a lexicalized feature that proved to be useful for pitch accent prediction [7]. 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.
- **Word identity**: Word identity refers to the word itself. This feature is motivated by the fact many words carried contrastive relations two or more times in the corpus.
- **Part-of-speech (POS)**: Broad POS with six broad categories (nouns, verbs, function words, pronouns, adjectives and adverbs) were used.
- **Only-same-POS**: If W1 is the only word in the sentence having the same broad POS as W2.
- **Closest-same-POS**: If W1 is the closest (in term of words between them) word preceding W2 and having the same broad POS as W2.
- **Textual similarity**: Score of two clauses containing (W1, W2). Since textual parallelism can be a clue of contrast, the parallelism (normalized) score was computed.
- **CAP relation**: If two-words are adverbial / prepositional phrases between (W1, W2), or one-word is adverbial / prepositional. This feature is used to capture contrastive relation triggered by "rather than", "or".
- **Suffix**: If one of the two words in the pair is contained within the other one (e.g. formal vs. informal).
- **Dependency relations**: Syntactic dependency relations involving (W1, W2) as dependents (e.g. subject-of).
- **Same dependency**: If W1 and W2 have the same type of dependency.
- **Same dependency head**: If W1 and W2 have the same type of dependency with their heads, and their heads refer to the same item. For example, in sentence "Is it doing a good job or a bad job?", both "good" and "bad" have a "modifier of" dependency with "job", and their heads refer to the same item "job".
- **WordNet**: Semantic relations (indicating if two words are linked by the relation: hypernyms, antonyms, entails, member-of, part-of, or sisters) was obtained using WordNet::QueryData [13] module. Semantic similarity was computed using the Word-Net::Similarity [13] module. Measures of similarity use information found in an is-a hierarchy of concepts (or synsets), and quantify how much concept A is like (or is similar to) concept B. For example, such a measure might show that an automobile is more like a boat than it is a tree, due to the fact that automobile and boat share vehicle as an ancestor in the WordNet noun hierarchy.

- **F0 Minimum, maximum, mean**: minimum, maximum, and mean f0 of each word in utterance.
- **Energy Maximum**: maximum energy of each word in utterance.
- **Duration**: Word duration extracted from the corpus.

2.3. Detecting Contrastive Word Pairs

Considering the limited amount of training data and the imbalance distribution between the positive and negative samples, SVMs [14] were used as the tagger for detecting contrastive word pairs from the above features. Specifically, we used LibSVM implementation [15], which has different kinds of kernels: linear, polynomial, radial basis, and sigmoid tanh. The training and testing set consisted of 3196 examples, 176 positive and 3020 negative. After trying different kernels, we found polynomial kernel with order 2 to be the best. The polynomial kernel with order higher than 2 seems to over-fit the data. Considering the unbalanced distribution between positive and negative examples, we set different costs on false positive and false negative; w_{-1} is defined as cost on false negative in LibSVM while w_1 is defined as cost on false positive. R in equation (1) measures the ratio between cost on false negative and cost on false positive. The tagger can achieve the best performance when R is set to 2.

$$R = \frac{w_{-1}}{w_1} \tag{1}$$

3. EXPERIMENTS

We conducted an objective experiment to evaluate the performance of the SVM tagger for automatic detection of contrastive word pairs and evaluate the importance of the newly proposed acoustic features.

Accuracy, precision and recall are used as the performance measure and defined as:

$$accuracy = \frac{TP + TN}{P + N} \tag{2}$$

$$precision = \frac{TP}{TP + FP} \tag{3}$$

$$recall = \frac{TP}{TP + FN} \tag{4}$$

Where P is the number all positive examples in training set, N is the number of all negative ones. TP is the number of positive examples correctly identified, TN is the number of negative examples correctly identified, FN is the number of positive examples incorrectly tagged as negative, and FP is the number of negative examples incorrectly tagged as positive.

Table II shows the performance of the tagger from 5-fold cross-validation using different features. The baseline is a tagger that always labels examples as non-contrastive, and gave 94.49% accuracy. The second row shows that by using all textual features, the accuracy increased to 95.06%, just as the result we had achieved in [7]. Adding the acoustic feature gives a big improvement up to 0.35%, and the accuracy turned out to be 95.41%. This suggested that contrastive words' acoustic characteristics are apparent different to non-contrastive words, and could be used to distinguish contrastive words from non-contrastive words.

Table II. Performance of contrastive word pairs detection with SVM using different features. R is the ratio between the cost on false negatives and the cost on false positives for SVM. The order of the polynomial kernel is 2. The baseline is a tagger that always labels examples as non-contrastive. After excluding acoustic features, the features are the same as used in [7].

Features	*R*	*Accuracy*	*Precision*	*Recall*
Baseline		94.49%	0	0
Textual	2	95.06%	64.06%	23.30%
Textual and Acoustic	2	95.41%	67.47%	31.82%

4. CONCLUSIONS

Labeling emphatic words from speech recordings plays an important role in building speech corpus for expressive speech synthesis. In this paper, we focused on the automatic detection of contrastive word pairs. For detecting the contrastive word pairs, we proposed to use SVMs as the tagger. We improved the accuracy of the tagger by adding acoustic features: F0 min, max, mean, energy max, duration. As in our analysis, contrastive words tend to have higher energy, F0 and longer duration, especially duration; therefore, these features are very helpful to distinguish contrastive words from non- contrastive words.

5. ACKNOWLEDGEMENTS

This work is partially supported by the National Natural Science Foundation of China (61375027, 60805008), the Major Project of National Social Science Foundation of China (13&ZD189), the NSFC/RGC Joint Research Scheme (N_CUHK 414/09), and Guangdong Provincial Science and Technology Program (2012A011100008).

REFERENCES

[1] E. Vallduvi and M. Vilkuna, “On rheme and kontrast,” *The Limits of Syntax (Syntax and Semantics 29)*, pp. 79-108, 1998.

[2] V. Strom, A. Nenkova, R. Clark, Y. Vazquez-alvarez, J. Brenier, S. King and D. Jurafsky, “Modeling prominence and emphasis improves unit-selection synthesis,” In: *Proc. Annual Conf. of Int. Speech Communication Association (Interspeech)*, 2007.

[3] J. Godfrey, E. Holliman and J. McDaniel, “Switchboard: Telephone speech corpus for research and development,” In: *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, pp. 517-520, 1992.

[4] T. Zhang, M. Hasegawa-Johnson and S.E. Levinson, “Extraction of pragmatic and semantic salience from spontaneous spoken English,” *Speech Communication*, vol. 48, pp. 437-462, 2006.

[5] V. Kumar, R. Sridhar, A. Nenkova, S. Narayanan and D. Jurafsky, “Detecting prominence in conversational speech: pitch accent, givenness and focus,” In: *Proc. Int. Conf. on Speech Prosody*, 2008.

[6] L. Badino and R. Clark, “Automatic labeling of contrastive word pairs from spontaneous spoken English,” In: *Proc. IEEE/ACL Workshop on Spoken Language Technology (SLT)*, pp. 101-104, 2008.

[7] C.R. Li, Z.Y. Wu, F.B. Meng, H. Meng and L.H. Cai, “Detection and emphatic realization of contrastive word pairs for expressive text-to-speech synthesis,” In: *Proc. Int. Symposium on Chinese Spoken Language Processing (ISCSLP)*, pp. 93-97, 2012.

[8] J. Francom and M. Hulden, “Parallel multi-theory annotations of syntactic structure,” In: *Proc. Int. Conf. on Language Resources and Evaluation (LREC)*, 2008.

[9] S. Calhoun, M. Nissim, M. Steedman and J. Brenier, “A framework for annotating information structure in discourse,” In: *Frontiers in Corpus Annotation II: Pie in the Sky, ACL Conference Workshop*, 2005.

[10] J. Terken and D. Hermes, “The perception of prosodic prominence,” *M. Horne (ed.): Prosody: Theory and experiment*, pp. 89-127, 2000.

[11] D.R. Ladd and R. Morton, “The perception of intonation emphasis: Continuous or categorical?” *Journal of Phonetics*, vol. 25, pp. 313-342, 1997.

[12] P. Boersma and D. Weenink, “Praat: doing phonetics by computer (Version 5.3.51),” [*Computer program*], retrieved 2 June 2013 from http://www.praat.org/.

[13] T. Pedersen, S. Patwardhan and J. Michelizzi, “WordNet::similarity - measuring the relatedness of concepts,” In: *Proc. Int. Conf. North American Chapter of Association for Computational Linguistics (NAACL)*, pp. 38-41, 2005

[14] T. Joachims, *Learning to Classify Text Using Support Vector Machines*, Kluver, 2002.

[15] C.C. Chang and C.J. Lin, “LIBSVM: a library for support vector machines,” *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, pp. 1-27, 2011.