DETECTION AND EMPHATIC REALIZATION OF CONTRASTIVE WORD PAIRS FOR EXPRESSIVE TEXT-TO-SPEECH SYNTHESIS

Chunrong Li [1,2], *Zhiyong Wu* [1,2,3], *Fanbo Meng* [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
qcclcr@gmail.com, zywu@sz.tsinghua.edu.cn, mfb03@mails.tsinghua.edu.cn,
hmmeng@se.cuhk.edu.hk, clh-dcs@tsinghua.edu.cn

ABSTRACT

This paper addresses the problem of automatic detection of contrastive word pairs and their acoustic realization in emphasis for expressive text-to-speech (TTS) synthesis in English. Support vector machines (SVMs) have been used to automatically detect contrastive word pairs from lexical features, syntactic dependencies and semantic relations. A much better performance is achieved by adding accent ratio and word identity features. Hidden Markov model (HMM) based speech synthesis is then used to generate emphatic speeches by putting emphasis on the detected contrastive word pairs. Subjective experiments show that most of the listeners consider putting emphasis on contrastive word pairs is more acceptable than on non-contrastive word pairs. This indicates the importance of the accurate detection of contrastive word pairs.

Index Terms— contrast, expressive text-to-speech (TTS) synthesis, support vector machines (SVMs), hidden Markov model (HMM) based speech synthesis

1. INTRODUCTION

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 more expressive and signaling the focus [1] of the sentence. Research [2] shows that appropriate emphasis improves the overall perception of synthesized speech as appropriate assignment of emphasis improves the expressivity and naturalness. In this study, we focus on the detection and synthesizing of contrastive word pairs for expressive TTS synthesis. The definition of contrastive word pairs is: an information structure relation that links two semantically related words that explicitly contrast with each other. a) and b) are examples of contrastive word pairs from the Switchboard corpus [3].

a) I think you could recover from a **pistol**, but not from a **gun**.
b) So well... **you** take this subject much more personally than **I** do, I suppose.

where "pistol" contrasts "gun", and "you" contrasts "I". To emphasize contrastive words, they can be synthesized with particular prominence than normal pitch accent (e.g. higher F0, longer duration, etc).

To generate contrastive emphasis, [4] marks contrast accent with prosodic ToBI labels, which are used to predict F0 and duration. Finally F0 and duration are used to select segments to minimize a cost function in unit selection. In [5], contrastive emphasis is realized with HMM-based speech synthesis. Speech is generated from a set of trained HMM models according to a sequence of context dependent labels including emphasis related labels, which were obtained from the text of test sentence.

Regarding the researches on automatic detection of contrastive word pairs, [6] 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 but distinct or contrastive in meaning. In [7], acoustic and lexical features are used to detect different classes of focus.

The latest and the most relevant to our work on automatic detection of contrastive word pairs is [8], in which 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 only features.

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 [8], and report the results in Section 2. In Section 3 we describe the approach and the features selected to train an HMM-based speech synthesis system, which can generate recognizable emphasis of contrastive words; subjective

experiment performed to measure the ability to convey emphasis is also described in this section. In the final section, we draw conclusions from our experiments and describe future directions for our research.

2. CONTRASTIVE WORD PAIRS DETECTION WITH SVM

We used support vector machines (SVMs) [14] as the tagger for automatic detection of contrastive word pairs from the lexical, syntactic and semantic features that are derived from the input text. In addition to the common features that were mentioned in [8], we also proposed two new features (accent ratio and word identity) for the task.

2.1. Data Set

Our experiment used a subset of the Switchboard corpus [3] that had been annotated with syntactic structure [9] and information structure [10]. 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, as shown in Table I. All word pairs sharing the same broad POS are extracted and then assigned a +1 (positive) if the word pair is linked with contrast or a -1 (negative) otherwise.

Table I. Contrast example value generation from sentence: Are/VBP they/PRP going/VBG to/TO be/VB taught/VBN nothing/NN or/CC they/PRP going/VBG be/VB taught/VBN something/NN. The contrastive word pair (nothing - something) 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).

W1	W2	Example value
are	going	-1
are	be	-1
are	taught	-1
…	…	…
nothing	something	+1
...	...	…

2.2. Features

2.2.1. Common features

The features considered for detection of contrastive word pairs include all features as mentioned in [8]. These features were text-based and could be grouped into three categories: lexical features, syntactic features and semantic features.

Examples of lexical feature are:

- 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 feature are:

- 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 features

Besides, we found that pitch accent and contrast were highly correlated. In [11], contrast seems to be helpful to predict pitch accent. Contrastive elements detection performance increases when it is performed only over the words that have been predicted to be accented [12].

We performed statistics of the distribution between the accented (bearing pitch accent) words and the contrastive categories for the sentences from the corpus, as shown in table II. As expected, pitch accent and contrastive status are highly correlated. Almost 43.39% of the accented words are contrastive while only 10.09% of the non-accented words are contrastive. In other words, the words bearing pitch accent are much more likely to be contrastive elements than the non-accented words.

A new feature called *accent ratio* is proposed, which is the estimated probability of the word being accented in a training corpus.

$$\text{Accent Ratio } (w) = \begin{cases} k/n & \text{if } B(k,n,0.5) \le 0.05 \\ 0.5 & \text{otherwise} \end{cases} \quad (1)$$

where k is the number of times word w appeared accented in the corpus, n is the total number of times the word w appeared, $B(k,n,0.5)$ is the probability (binomial distribution) that k successes occur out of n trials. Accent ratio was computed over 75 Switchboard conversations annotated with pitch accent.

Table II. Corpus distribution statistics across accented and contrastive categories. The two are highly correlated, words bearing accent tend to be more possible to be contrastive elements than words not bearing accent.

	Contrastive	Non contrastive
Accented	1778 (43.39%)	2320 (56.61%)
Non-accented	372 (10.09%)	3315 (89.91%)

Moreover, we found that some words are more likely to be contrastive elements, as shown in table III. Up to 30% of all contrastive words in the corpus carried contrastive relation two or more times. For example "men" and "women"

occurred to be contrastive word pair 5 times, while "good" and "bad" word pair occurred 4 times. Hence, another new feature called *word identity* is proposed, which refers to the English word itself.

Table III. Occurring times of words bearing contrastive relation in the corpus, some words are more likely to be contrastive elements than others.

W1	W2	Occurring times
Men	Women	5
Good	Bad	4
Here	There	4
More	Less	3
Buy	Sell	2

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.

- **Part-of-speech**: 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.
- **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".
- **Textual similarity**: Score of two clauses containing (W1, W2). Since textual parallelism can be a clue of contrast, the parallelism (normalized) score was computed.
- **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.
- **Suffix**: If one of the two words in the pair is contained within the other one (e.g. formal vs. informal).
- **Accent ratio**: Probability of the word being accented (bearing pitch accent) in corpus 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.

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 (2) 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.

$$R = w_{-1} / w_1 \tag{2}$$

3. EMPHATIC REALIZATION WITH HMM-BASED SPEECH SYNTHESIS

HMM-based synthesis (HTS) [16] provides a data-driven framework that allows finer-grained control of expressivity of speech, by learning models mapping context features to individual speech parameters. Recent work shows that HMM-based synthesis can produce recognizable variation when modeling emphasis of contrastive words [17].

3.1. Speech Data

To produce exaggerated emphasis that signaling the focus of sentence to draw interlocutor's attention, 350 text prompts were carefully designed [18]. Three example text prompts are as follows (with focus word in italic boldface):

- "On the area of the ***sea***, the pandas like to drink tea with ***peas*** in soda."
- "A ***mistake*** in ***staking*** can overtake you like an earthquake."
- "A ***mule*** of ***molecule*** has been ***scheduled***."

Two contrastive utterances were recorded for each text prompt, one with expressive intonation to put emphasis on focus words and the other with neutral intonation throughout the utterance. 700 utterances were recorded by a female speaker with high level of English proficiency.

3.2. Decision Tree Clustering with Additional Emphasis-

related Context Questions

In the training stage, to deal with the issue of limited amount of training data, the context clustering is used. Information sharing of training data in the same cluster (or the leaf node in the decision tree based context clustering) is the essential concept. To construct decision tree for generating emphasis, context questions must be carefully designed. In our work, a set of emphasis-related questions are designed, as shown in Table IV, in addition to the standard context questions (non-emphasis related) from the official HTS toolkit [16]. In our work the focus word was set manually (on either contrastive or non-contrastive word pairs) to evaluate the preference of the subjects (listeners) for the synthesized speech.

Table IV. Emphasis-related questions and related answers for building decision tree to generate emphasis.

Emphasis-related Questions	Answer
Is the phone in stressed syllable of focus word	0/1
Is the phone before stressed syllable of focus word	0/1
Is the phone after stressed syllable of focus word	0/1
Is the phone in the neutral word before focus word	0/1
Is the phone in the neutral word after focus word	0/1
Other situation	0/1

4. EXPERIMENTS

4.1. Contrastive Word Pairs Detection Experiment

We first 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 features (accent ratio and word identity).

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

$$accuracy = (TP + TN)/(P + N) \quad (3)$$

$$precision = TP/(TP + FP) \quad (4)$$

$$recall = TP/(TP + FN) \quad (5)$$

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, FN is the number of positive examples incorrectly tagged as negative, FP is the number of negative examples incorrectly tagged as positive.

Table V 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 the features without accent ratio and word identity (i.e. features in [8]), the accuracy increased to 94.87%, which was comparable with the result in [8]. Adding the accent ratio feature gave a further improvement up to 0.12%, and the accuracy turned out to be 94.99%. This suggested that detecting contrast among the pith-accent words based on accent ratio could improve the performance. Finally by adding word identity, we got the final accuracy of 95.06%. Word identity turned out to be highly ranked as a feature in the system. This could be attributed to the fact that words occurring two or more times in the corpus carrying an emphatic pitch accent accounted for 30% of emphatic tokens in the corpus.

Table V. 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 accent ratio and word identity, the features are the same as used in [8].

Features	*R*	*Accuracy*	*Precision*	*Recall*
Baseline		94.49%	0	0
All features without Accent ratio and Word identity	2	94.87%	56.25%	30.68%
All features without Word Identity	2	94.99%	61.11%	25.00%
All features	2	95.06%	64.06%	23.30%

4.2. Subjective Experiments on Emphasis Realization

Two experiments were conducted to evaluate the realization of emphasis for contrastive word pairs by listening test. The first experiment was to evaluate if the proposed method could generate appropriate emphatic speech. And the second experiment was to validate the importance of correct detection of contrastive word pairs by preference test.

4.2.1. Experiment on emphasis intensity

This experiment was conducted to evaluate if the methods can properly generate emphatic speech. We selected 20 sentences containing one contrastive word pair and synthesized them with emphasis on a single word (any one word in the contrastive pair). After listening, subjects were instructed to select the most prominent word they perceived in the utterance. 13 subjects participated in the experiment. On average, subjects were able to detect the emphatic words in 16 out of 20 sentences.

4.2.2. Experiment on preference of emphasis on word pairs

10 sentences whose contrastive word pairs were correctly detected by our SVM tagger were selected from the whole dataset, and were synthesized with different conditions:

- A: synthesized with no emphasis.
- B: synthesized with emphasis on contrastive words.
- C: synthesized with emphasis on non-contrastive words having the same broad POS as contrastive words.

Overall results on figure 1 show that 35% listeners preferred the utterances with emphasis on contrastive words (B), while only 10% listeners preferred sentences with emphasis on non-contrastive words (C). Listeners had an apparent preference to the former, which suggested that non-contrastive words were less suitable to carry emphasis, and that contrastive word pairs must be correctly identified.

Hence, it is important to improve the detection performance of the contrastive word pairs. However, it can also be seen that 55% of the listeners had a preference for utterances without emphasis (A). This may be due to the reason the emphasis generated by our method was sometimes too strong to degrade the naturalness of the synthetic sentence, which further affected the listeners' preference choice.

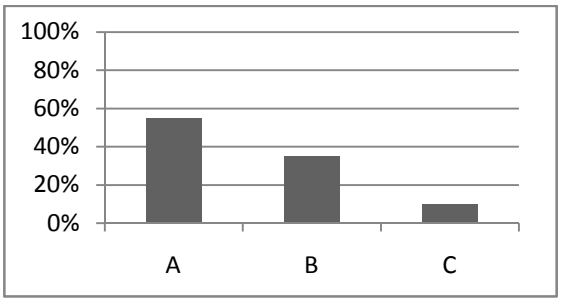

Fig.1. The result of the preference experiment by putting emphasis on different word pairs. Sentences were synthesized under A (no emphasis), B (putting emphasis on contrastive word pairs) and C (putting emphasis on non-contrastive words) conditions.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we focused on the automatic detection of contrastive word pairs and their acoustic realization in emphasis for expressive TTS synthesis. For detecting the contrastive word pairs, we proposed to use SVMs as the tagger. We improved the accuracy of the tagger by adding two effective features: accent ratio and word identity. As in our analysis, accented words are more likely to be contrastive than non-accented words; and lots of the words in the corpus carried contrastive relation more than one time. As for the acoustic realization of contrastive word pairs, we proposed to use HMM-based speech synthesis to generate emphasis for the contrastive word pairs. Subjective experiment shows that non-contrastive word pairs are less appropriate to carry emphasis than contrastive words. It indicates that improving the performance of contrast tagger as we have done is very important. However, too strong emphasis affects the naturalness and acceptability of the synthesized speech, new methods for emphatic speech synthesis should be investigated in the future to achieve good performance in both naturalness and emphasis. Furthermore, emphasis in sentences generally signals new or important information and also ends of intonation phrases; but sometimes it may change semantics or focus of the sentences, and affects listener's preference as a result. We will conduct experiments to evaluate this in our future work.

6. ACKNOWLEDGEMENTS

This work is partially supported by the National Natural Science Foundation of China (60928005, 60805008, 60931160443 and 61003094) and the NSFC/RGC Joint Research Scheme (N_CUHK 414/09).

7. REFERENCES

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

[2] V. Strom, A. Nenkova, R. Clark, Y. Vazquez-Alvarez, J. Brenier, S. King, and D. Jurafsky, "Modelling prominence and emphasis improves unit-selection synthesis," In: *Proc. Interspeech*, 2007.

[3] J. Godfrey, E. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," In: *Proc. ICASSP*, 1992.

[4] J. Pitrelli, and E. Eide, "Expressive speech synthesis using American English tobi: Questions and contrastive emphasis," In: *Proc. IEEE ASRU*, 2003.

[5] K.Yu, F. Mairesse, and S. Young, "Word-level emphasis modeling in HMM-based speech synthesis," In: *Proc. ICASSP*, pp. 4238-4241, 2010.

[6] 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.

[7] V. Sridhar, A. Nenkova, S. Narayanan, and D. Jurafsky, "Detecting prominence in conversational speech: pitch accent, givenness and focus," In: *Proc. Speech Prosody*, 2008.

[8] L. Badino, and R. Clark, "Automatic labeling of contrastive word pairs from spontaneous spoken English," In: *Proc. SLT*, 2008.

[9] J. Francom, and M. Hulden, "Parallel multi theory annotations of syntactic structure," In: *Proc. LREC*, 2008.

[10] 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, ACL2005 Conference Workshop*, 2005.

[11] A. Nenkova, J. Brenier, A. Kothari, S. Calhoun, L. Whitton, D. Beaver, and D. Jurafsky, "To memorize or to predict: Prominence labeling in conversational speech," In: *Prof. NAACL-HLT*, 2007.

[12] A. Nenkova, and D. Jurafsky, "Automatic detection of contrastive elements in spontaneous speech," In: *Proc. IEEE ASRU*, 2007.

[13] T. Pedersen, S. Patwardhan, and J. Michelizzi, "Wordnet::similarity - measuring the relatedness of concepts," In: *Proc. NAACL*, 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*, 2:27:1-27, 2011.

[16] K. Tokuda, H. Zen, and A. Black, "An HMM-based speech synthesis system applied to english," In: *Proc. IEEE SSW*, 2002.

[17] L. Badino, J. S. Andersson, J. Yamagishi, and R. Clark, "Identification of contrast and its emphatic realization in hmm based speech synthesis," In: *Proc. Interspeech*, 2009.

[18] F.B. Meng, Z.Y. Wu, H. Meng, J. Jia, and L.H. Cai, "Hierarchical English emphatic speech synthesis based on HMM with limited training data," In: *Proc. Interspeech*, 2012.