INTERSPEECH 2009 BRIGHTON

Identification of Contrast and Its Emphatic Realization in HMM based Speech Synthesis

Leonardo Badino, J. Sebastian Andersson, Junichi Yamagishi, Robert A.J. Clark

Centre for Speech Technology Research
University of Edinburgh, Edinburgh, UK
l.badino@sms.ed.ac.uk

Abstract

The work presented in this paper proposes to identify *contrast* in the form of contrastive word pairs and prosodically signal it with emphatic accents in a Text-to-Speech (TTS) application using a Hidden-Markov-Model (HMM) based speech synthesis system.

We first describe a novel method to automatically detect contrastive word pairs using textual features only and report its performance on a corpus of spontaneous conversations in English. Subsequently we describe the set of features selected to train a HMM-based speech synthesis system and attempting to properly control prosodic prominence (including emphasis).

Results from a large scale perceptual test show that in the majority of cases listeners judge emphatic contrastive word pairs as acceptable as their non-emphatic counterpart, while emphasis on non-contrastive pairs is almost never acceptable.
Index Terms: prosody, contrast, hmm speech synthesis

1. Introduction

The work presented here aims to improve prosodic expressiveness and realization of context-dependent prosody in TTS synthesis by automatically identifying contrast activated by pairs of contrastive words and prosodically marking this contrast with emphatic pitch accents. The following is an example of a contrastive word pair from the Switchboard corpus [1]:

(1) I've never really joined a club because I haven't got the **time**. Not because I haven't got the **desire**.

Contrastive *focus*, meant more generally as focus evoking a restricted set of alternatives explicitly given by the discourse context (e.g. "*Helen, Paul and Martin came to my party. Only* ***Paul*** *brought a present.*"), has been claimed to be signaled by a particularly prominent pitch accent(see [2] for example). It is not clear, however, whether this particular prominence is due to intrinsic properties of the contrastive pitch accent (e.g. higher F0 peak) or is relative to the prosodic context, meaning that when heard out of prosodic context it is not perceived as more prominent than "ordinary" accent [3].

In this work we decided to use emphatic speech already available from the Blizzard Challenge [4] speech corpora, to prosodically mark contrastive word pairs in order to increase the expressiveness of our synthetic voice according to the semantic, pragmatic and syntactic context. Obviously it is ultimately up to the speaker (and their intentional state) whether to emphasise contrastive words or not, but from a TTS perspective, contrastive words are justifiable candidates for emphasis while words that are not identifiable as contrastive from their textual context are not candidates to carry emphatic accents, unless we want to take the risk of generating prosodic prominent patterns that collide with the discourse context.

There is little previous works concerned with the generation of contextually appropriate prosody in TTS synthesis. Perhaps the closest previous work is [5] in which contrastive accents were modeled with ToBI labels. The ToBI labels were used to predict F0 and duration which in turn were used as specification features in the cost function of a unit selection TTS system. Similar approaches were used in [6] and [7] to model (*thematic* and *rhematic*) focus in limited domain speech synthesis.

The main differences between previous work and ours resides in the speech synthesis techniques used and the fact that here we automatically detected concepts relating to discourse context (i.e. *contrast*) that are prosodically signaled in human speech and we do not rely on information given by a dialogue system or by the speech synthesis user through a mark-up language.

Looking for previous work in which discourse-level information is automatically extracted for TTS purposes we have to go back to [8] and [9]. For example, [8] automatically identifies new, given (i.e. already mentioned in the discourse) and "topical" (i.e. "[belonging] to concepts central to the main purpose of the discourse") words to improve pitch accents prediction for TTS synthesis.

Concerning the automatic detection of contrastive focus, [10] propose a combined use of acoustic features (e.g. F0, spectral balance cepstral coefficients), Part-Of-Speech (POS), and a semantic similarity measure (computed by using both the WordNet semantic lexicon and corpora statistics) to automatically label *symmetric contrast*, a scenario of contrastive focus which "consists of a set of words that are parallel or symmetric in linguistic structure but mutually exclusive in meaning". In [11] a subsection of the Switchboard corpus annotated by [12] is used to detect different sub-categories of contrastive focus. One of these categories is exactly the same category we try to identify in this work, i.e. contrastive focus activated by contrastive word pairs. The method proposed by [11] looks at acoustic properties and POS.

Relying on POS is obviously not enough to identify a contrastive relation between two words and much more complex textual features are necessary as we proposed in [13].

In the next section we describe the *contrast* tagger proposed in [13] and some modifications we have made in order to make it usable in TTS synthesis. Subsequently we describe the set of features selected to train an HMM-based speech synthesis system that could accurately control prosodic prominence. We conclude the paper by presenting the results of a perceptual experiment in which listeners where asked to judge the emphatic realization of contrastive words vs. a non-emphatic realization.

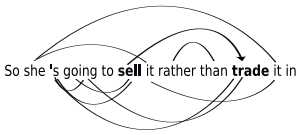

Figure 1: *Example values generation. The contrastive word pair (sell-trade) is given value +1. All the other possible pairs of words sharing same broad POS are given value -1.*

2. Contrast detection

The contrast tagger used in this work is an up-to-date version of the tagger we proposed in [13]. In [13] we selected all the examples of salient contrastive word pairs whose annotation is described in [12]. We only looked at contrastive word pairs occurring within the same sentence.

For each sentence both positive and negative examples of *contrast* were extracted as shown in Fig.1. All word pairs sharing the same broad POS were extracted and then assigned a +1 if the two words were linked by *contrast* or a -1 otherwise. Each example used to train or test the *contrast* tagger consisted of its positive or negative value and a sequence of training features.

For each example we extracted hundreds of features which can be roughly grouped into three categories: lexical, syntactic and semantic. Most of the features were intended to capture the fact that contrastive words are "comparable but dissimilar". Examples of lexical features are:

- all two-word adverbial and prepositional phrases between (w1,w2)
- textual similarity score of the two clauses containing (w1,w2).

Examples of syntactic features are:

- dependency relations (DR) involving (w1,w2) as dependents (e.g. subject-of)
- is (DR(head1,w1) = DR(head2,w2)) AND (head1=head2)?

Semantic features are mainly based on semantic relations extracted from the WordNet semantic lexicon (e.g. antonymy, member-of, etc...)

The main difference between the tagger presented in [13] and its new version is that instead of relying on gold syntactic dependencies the new version uses dependencies automatically extracted using the Malt dependency parser [14]. As a consequence all training features used by our tagger are all automatically extracted from text, making the tagger usable within a TTS system. With respect to the data used in [13] some sentences containing contrastive word pairs were removed because MaltParser splits the sentences into two or more sentences such that the contrastive words did no longer belong to the same sentence anymore.

Other minor changes were made, consisting of a couple of small bug fixes, some new features created from the conjunction of old ones, and the introduction of a morphological feature indicating if one of the two words in the pair is contained within the other one (e.g. formal vs. *in*formal) and the identity of the morpheme that differentiates them.

Finally the new version was trained on a new data set consisting of 246 positive examples and 7405 negative examples extracted from 220 sentences. The tagger is a Support Vector Machines based tagger. It was evaluated with a "stratified" cross-validation in which all the examples contained in one single sentence were held out, i.e. we carried out a leave-one(-sentence)-out procedure. Results are shown in Table 1.

	Accuracy	Precision	Recall
Baseline	96.8%	0%	0%
Tagger	97.3%	74.1%	24.4%

Table 1: Leave-one-out (on sentences) evaluation of the fully automatic *contrast* tagger. The baseline is a majority baseline assigning -1 to all examples. Precision and recall are relative to positive examples.

The high precision rate in the labeling of contrast makes the tagger reliable for TTS applications. However the tagger still has poor recall. The main reasons for such poor recall are:

- the presence of several false negative examples in the data.
- the limited number of positive examples compared with the possible scenarios triggering contrast between two words.
- the training features used are insufficient. However we believe that the additional training features needed for this task can not obtained from present Natural language processing tools.

We are currently working to address the first two points above, for example, by using *Active Learning* techniques to increase the number of training examples.

3. Realising Emphasis in HMM-based Speech Synthesis

HMM-based speech synthesis is currently a very active research area and a lot of progress has been made since it was first introduced by [15].

Building a speaker dependent HMM-based voice consists of extracting acoustic parameters from a speech database and training a set of context dependent HMMs. The context dependent specifications are obtained from textual analysis based on the text of speech speech and consist of both phonetic and linguistic information. The use of both phonetic and linguistic information results in a very large set of models that are clustered with decision tree-based clustering. At synthesis time speech is generated from trained models according to a sequence of context dependent labels obtained from the text of a test sentence [16].

In this work an HMM-based voice was built with the system configurations described in [16] where we changed the specifications of the context dependent labels to allow for synthesis of emphasis.

3.1. Prosody as Context Dependent Phonemes

The context dependent phonemes determine the phonetic, linguistic and prosodic categories for training as well as generation. The specifications of context dependent phonemes for neutral English speech is generally very similar to [17], where prosodic prominence is restricted to lexical stress and pitch accents, and most contexts are counts, positions and distances of phonemes, syllables, words and phrases.

In the speech synthesis Blizzard Challenge 2008 [4] some teams included emphasis contexts in HMM-based speech synthesis systems, but no results were reported, and our work is the first evaluation of emphasis realizations in an HMM-based voice.

As part of an ongoing investigation into prosodic modelling through context dependent phonemes in HMM-based speech synthesis we selected a different set of contexts than [17] on the basis that there was potentially important information missing, and too many contexts had rather opaque prosodic relevance.

Instead we selected a minimal set of concrete contexts within a more controllable prosodic window of at most preceding, current and succeeding word:

- *which* {preceding, current, succeeding} phoneme (e.g. uh1)
- *which* {preceding, current, succeeding} syllable (e.g. b_uh1_t)
- *which* {preceding, current, succeeding} word (e.g. but)

The phoneme and syllable names both included lexical stress (0,1,2). Phonemes were clustered both on articulatory features and stress level. For the word context, clustering was only applied to words with frequency above 20 in the training data, which limited the word context to mainly closed class words, and thereby separated function from content words. A distinction was made between utterance internal and beginning/final silences.

The contexts for pitch accent and emphasis were binary values set for pitch accents on:

- *which* current syllable nucleus
- {preceding,current,succeeding} syllable
- {preceding,current,succeeding} word

And for emphasis on:

- *which* {preceding, current, succeeding} phoneme
- *which* current syllable nucleus
- {preceding, current, succeeding} syllable

Pitch accents and emphasis did not have the same context specification, this was partly motivated by emphasis being stronger than pitch accent and hence affects nearby phones more, and partly by all our emphasis being in carrier sentences (see section 3.2) where a larger prosodic window might have resulted in modelling artefacts.

Pitch accents were automatically predicted using a slightly improved version of the accent predictor described in [18]. The accuracy of the predictor is around 85% when trained and tested on read news speech.

3.2. Speech Data

The speech was selected from a database that had previously been used to synthesise emphatic accents with unit selection and so had existing mark-up of emphasis [19]. From this database we used *Arctic* containing 1132 utterances for general phonetic coverage, and 1683 carrier sentences containing more than 1100 emphasised names in the following template format:

"It was JAMES who did it."

"No, it was JOHN who did it."

"It was JOHN, not JAMES!"

3.3. Resulting Voice

Informal listening tests of the resulting voice suggested that the general quality was good, and that pitch accents made a positive impact on the quality of the speech and that emphasis could be realized.

4. Listening test

We conducted a formal listening experiment to investigate whether emphatic realizations of contrastive word pairs detected by the contrast tagger were at least as acceptable as non-emphatic counterparts. We also generated utterances in which pairs of non-contrastive words were emphasised in order to see if emphasis on non-contrastive words were equally acceptable. Note that emphasis on non-contrastive words may be acceptable since emphasis is not only used to mark a contrastive relation between two words (e.g. an out-of-context utterance like "That man went to **Madrid** by **bike**!".

4.1. Test design

Test subjects were asked to carry out two tasks: the first to test the acceptability of emphasis on word pairs and the second to test how well listeners recognized intended emphasis.

Test 1. We selected 20 sentences from the whole set of sentences where our contrast tagger correctly identified contrastive word pairs. The selection was carried out trying to have as many diverse scenarios of *contrast* as possible. So, for example, we included contrast triggered by comparison (e.g. "**They** have probably had more time than **you** had to think about this subject") or by antonymy (e.g. "Every time we get a real **good** player they treat him **bad**."). All 20 sentences were synthesised with emphasis on both contrastive words and without any emphasis. A subset of 10 sentences was synthesised with emphasis on two non-contrastive words having the same POS as the identified contrastive pair which were synthesised with no emphasis. Note that one of the two emphasised words could belong to the contrastive pair. In the end we synthesised 30 (20 "contrast" + 10 "no-contrast") sentence pairs (emphatic vs. non-emphatic), which were presented in both internal orders, so each participant listened to 60 utterance pairs in a randomized order. For each pair, the participants could express a preference for one of the two utterances, or no preference.

Test 2. We selected 14 sentences containing at least one contrastive word pair and synthesised them with emphasis on a single word (that could be a word in the contrastive word pair). Participants were instructed to select the word they perceived as most prominent. 36 subjects were recruited, all of them were native English speakers. The tests were conducted through a web browser and lasted approximately 30 minutes each.

4.2. Results

Test 1. Overall results show that listeners had a significant preference for utterances without emphasis (see Table 2). However, when looking at each of the 20 "contrast" pairs the difference between emphatic vs. non-emphatic utterances is less evident with only 8 out of 20 pairs in which listeners had a significant ($p < 0.05$) preference for the non-emphatic utterances.

Of the 10 "no-contrast" pairs, subjects significantly preferred the non-emphatic utterances in 9 out of 10 sentences. This supports our hypothesis that listeners prefer emphasis on contrastive word pairs than on non-contrastive pairs showing that contrastive pairs are better candidates for emphasis. How-

	Emph.	No-preference	NEmph.	p-value
Contrast	299(21%)	461(32%)	680(47%)	$p < 0.00001$
No-contrast	110(15%)	180(25%)	430(60%)	$p < 0.00001$

Table 2: *Emphatic vs. Non-emphatic utterances. Contrast* refers to the 20 utterance pairs in which the emphatic utterance had emphasis on the contrastive words. *No-contrast* refers to the 10 pairs in which the emphatic utterances had emphasis on non-contrastive word pairs. Emph. refers to number of times participants preferred emphatic utterances, NEmph to the number of times they preferred non-emphatic utterances. Columns 2-4 report the number of preferences for the three options. The p-values are from two-sided Binomial tests and were computed after summing two equal halves of *No-preference* to *Emph.* and *NEmph.*.

ever the overall preference for non-emphatic realizations is undoubtedly disappointing. We believe that this was mainly due to the generated emphasis being too strong and so often inappropriate. When emphasis was not too strong, for example, when emphasis was on pronouns (e.g. "**They** have probably had more time than **you** had to think about this subject"), participants had no clear preference.

Test 2. Results of Test 2 shows that we were able to generate detectable emphasis. In 12 out of 14 sentences the number of speakers able to detect the emphatic word was significantly much higher than chance level, with $p - value \ll 0.01$ on a two-sided binomial test and only accented and emphatic words taken into account to compute the chance level.

To realise emphasis perceived as less strong we could: 1) normalise the two speech databases with respect to each other; 2) compand the synthesised sentences to reduce the dynamic variability; 3) apply further rules as to when it is appropriate to use strong emphasis; 4) or the ideal but difficult solution: record data with emphatic accents within natural sentences from a desired text genre.

5. Conclusion

The work presented in this paper aimed to identify *contrast* in the form of contrastive word pairs and prosodically signal it with emphatic accents. We described a novel method to automatically identify *contrast* using only textual features. We then described a possible way to build a HMM-based voice capable of realizing emphasis. A large scale listening test showed that contrastive word pairs are more appropriate to emphasise than non-contrastive words. However the realization of emphasis turned out to be occasionally strong and therefore less contextually appropriate than "standard" pitch accents.

6. Acknowledgements

2nd author is supported by by Marie Curie Early Stage Training Site EdSST (MEST-CT-2005-020568).

This work has made use of the resources provided by the Edinburgh Compute and Data Facility (ECDF). (http://www.ecdf.ed.ac.uk/).

7. References

[1] J. Godfrey, E. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *Proceedings of ICASSP 92*, 1992.

[2] J. Pierrehumbert and J. Hirschberg, *Intentions in Communication*. Cambridge MA.: MIT Press, 1990, ch. The Meaning of Intonation in the Interpretation of Discourse, pp. 271–311.

[3] E. Krahmer and M. Swerts, "On the alleged existence of contrastive accents," *Speech Communication*, vol. 34, no. 4, 2001.

[4] V. Karaiskos, S. King, R. Clark, and C. Mayo, "The blizzard challenge 2008," in *The Blizzard Challenge*, Brisbane, Australia, 2008.

[5] J. Pitrelli and E. Eide, "Expressive speech synthesis using american english tobi: Questions and contrastive emphasis," in *Proceedings of IEEE ASRU*, St. Thomas, U.S. Virgin Islands, December 2003.

[6] S. Prevost and M. Steedman, "Specifying intonation from context for speech synthesis," *Speech Communication*, vol. 15, pp. 139–153, 1994.

[7] R. Baker, R. Clark, and M. White, "Synthesising contextually appropriate intonation in limited domains," in *Proc. 5th ISCA workshop on speech synthesis*, Pittsburgh, USA, 2004.

[8] J. Hirschberg, "Pitch accent in context: Predicting intonational prominence from text," *Artificial Intelligence*, vol. 63, pp. 305–340, 1993.

[9] L. Hiyakumoto, S. Prevost, and J. Cassell, "Semantic and discourse information for text-to-speech intonation," in *Proceedings of the ACL Workshop on Concept to Speech Generation Systems*, 1997, pp. 47–56.

[10] 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.

[11] V. Sridhar, A. Nenkova, S. Narayanan, and D. Jurafsky, "Detecting prominence in conversational speech: pitch accent, givenness and focus," in *4th Conference on Speech Prosody*, Campinas, Brazil, 2008.

[12] S. Calhoun, S. Nissim, M. Steedman, and J. Brenier, "A framework for annotating information structure in discourse," in *Frontiers of Corpus Annotation II: pie in the Sky, ACL 2005 Conference Workshop*, Ann Arbor, Michigan, 2005.

[13] L. Badino and R. Clark, "Automatic labeling of contrastive word pairs from spontaneous spoken english," in *2008 IEEE/ACL Workshop on Spoken Language Technology*, Goa, India, 2008.

[14] J. Nivre, J. Hall, J. Nilsson, A. Chanev, G. Eryigit, S. Kbler, S. Marinov, and E. Marsi, "Maltparser: A language-independent system for data-driven dependency parsing," *Natural Language Engineering*, vol. 13, no. 2, pp. 95–135, 2007.

[15] K. Tokuda, T. Kobayashi, and S. Imai, "Speech parameter generation from HMM using dynamic features," in *Proc. ICASSP*, 1995.

[16] H. Zen, T. Toda, M. Nakamura, and K. Tokuda, "Details of the Nitech HMM-based speech synthesis system for the blizzard challenge 2005," *IEICE Transactions on Information and Systems*, vol. E90-D, no. 1, 2007.

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

[18] L. Badino, R. Clark, and V. Strom, "Including pitch accent optionality in unit selection text-to-speech synthesis," in *In Proc. Interspeech*, Brisbane, 2008, 2008.

[19] 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 *Interspeech*, Antwerp, Belgium, 2007, pp. 1282–1285.