Dialogue Context Sensitive Speech Synthesis using Factorized Decision Trees

Pirros Tsiakoulis, Catherine Breslin, Milica Gašić, Matthew Henderson, Dongho Kim, Steve Young

University of Cambridge, Engineering Department, Cambridge, UK

pt344@cam.ac.uk

Abstract

This paper extends our recent work on rich context utilization for expressive speech synthesis in spoken dialogue systems in which significant improvements to the appropriateness of HMM-based synthetic voices were achieved by introducing dialogue context into the decision tree state clustering stage. Continuing in this direction, this paper investigates the performance of dialogue context-sensitive voices in different domains. The Context Adaptive Training with Factorized Decision trees (FD-CAT) approach was used to train a dialogue context-sensitive synthetic voice which was then compared to a baseline system using the standard decision tree approach. Preference-based listening tests were conducted for two different domains. The first domain concerned restaurant information and had significant coverage in the training data, while the second dealing with appointment bookings had minimal coverage in the training data. No significant preference was found for any of the voices when tested in the restaurant domain whereas in the appointment booking domain, listeners showed a statistically significant preference for the adaptively trained voice.

Index Terms: HMM-based expressive speech synthesis, dialogue context-sensitive speech synthesis, context adaptive training, factorized decision trees

1. Introduction

Spoken Dialogue Systems (SDS) offer the potential for more natural human-machine interaction by using speech as the primary modality [1]. However, public acceptability of spoken interfaces is still limited by weaknesses in the technology. Apart from the challenges of robust recognition and dialogue management, the quality of speech generation remains a critical factor, since this is what the user perceives most directly.

Traditionally, a general purpose synthetic voice with neutral characteristics has been used for such applications. Recent effort has focused on making the discourse more natural by incorporating spontaneous responses, backchannel and fillers, as well as incremental processing [2, 3, 4, 5, 6, 7]. Expressive speech synthesis that is aware of the discourse context is an essential requirement in natural conversational dialogue. HMM-based speech synthesis has recently become a popular paradigm for expressive speech synthesis providing high quality synthetic speech and offering flexible control over both the acoustic and prosodic cues of the speech signal [8]. Moreover, adaptation techniques are inherently supported by the HMM approach enabling a variety of emphatic effects to be incorporated.

In a recent study, we have shown that a voice trained with dialogue context questions included in the decision tree state clustering was significantly preferred over the baseline voices for dialogue applications [9]. More specifically, an expressive dialogue corpus was collected for the tourist information domain (TownInfo), designed to include a rich set of styles appropriate for dialogue prompts including emphasized slots. Using this expressive dialogue corpus (in conjunction with a general purpose text-to-speech corpus recorded by the same speaker) an HMM-based voice was trained using emphasis context features, as well as context features extracted from the dialogue act semantic representation. A listening experiment designed for dialogue appropriateness in the restaurant domain showed that the dialogue context-sensitive voice was significantly preferred over baseline alternatives that were trained on the same data but did not include dialogue context features.

An expressive dialogue corpus consisting of prompts for a specific domain is inevitably phonetically unbalanced. Consequently, the use of a simple decision tree for this data leads to unbalanced modeling of the contextual factors, e.g. the phoneme sequence of a specific slot value that appears emphasized several times can be modeled adequately, while other sequences of phonemes may lack emphatic context and cannot be modeled as such. The context-sensitive voice may still be preferred for the in-domain dialogue task, however it may not extrapolate well if applied to a different domain.

In this paper, we investigate this issue by evaluating the performance of our dialogue context-sensitive TownInfo voice in a significantly different dialogue domain. For the latter, the appointment booking domain was selected since this is an application for which we have a working dialogue system but which has little overlap with the TownInfo response generator and hence minimal coverage in the domain specific training data. Initial experiments using standard decision tree clustering suggested that the unbalanced training data does indeed lead to degradation in the out-of-domain task. Hence, in an attempt to mitigate this, a voice was also built using Context Adaptive Training with Factorized Decision trees (FD-CAT) [10].

The remainder of this paper is structured as follows. Section 2 reviews related work in integrating dialogue context information with text to speech in order to generate contextually appropriate system responses. Section 3 then briefly reviews the FD-CAT approach to HMM synthesis. Sections 4 and 5 then describe the data used, the experimental procedure and the results. Section 6 presents conclusions.

2. Contextually-sensitive TTS

The term *Concept-To-Speech* (CTS) is used to describe methods that combine joint natural language generation (NLG) and text-to-speech (TTS) functionality, i.e. using semantic information as input to the speech synthesizer [11]. One approach to CTS involves an annotation schema which is applied to the generated text and affects the prosody of the rendered speech [12]. A sim-

The research leading to this work was funded by the EC FP7 programme FP7/2011-14 under grant agreement no. 287615 (PARLANCE).

ilar technique applies prosodic annotations to a template-slot based generation system [13]. Another approach is to jointly optimize text and prosody generation in the framework of unit selection TTS [14, 15]. Others have focused on prosody models driven from semantic as well as linguistic input [16, 17, 18].

The HMM based statistical speech synthesis framework (HTS) facilitates data-driven approaches for building voices offering high quality synthetic speech in addition to flexible control over both the acoustic and prosodic cues of the speech signal [8]. HTS uses decision trees to cluster and model the acoustic-prosodic space. The decision trees are built in a data-driven manner using linguistic information extracted from text. Any paralinguistic or non-linguistic information can be used as long as it can be predicted from text or input otherwise. Adaptive training and acoustic factorization originally introduced for acoustic modeling in the context of automatic speech recognition (ASR) have been recently applied for a variety of applications such as voice morphing, multi- and cross lingual speech synthesis, emotional speech synthesis and style control, etc. [19, 10, 20, 21].

Several efforts for modeling emphasis have been proposed in the framework of HMM-based speech synthesis. In most cases, a data-driven approach is followed, either by detecting/annotating emphasized words in existing corpora [22, 23] or by collecting speech corpora specifically designed for emphasis modeling [24]. Emphasis context features are then used for decision tree state clustering. More elaborate techniques have also been proposed that can tackle data sparsity issues when the emphasis data is limited, such as factorized decision trees [23, 10], hierarchical modeling [25], and phrase level modeling [26].

In our recent work, the HTS framework was successfully utilized to introduce dialogue context information in addition to emphasis in order to train a synthetic voice tailored to the needs of spoken dialogue systems [9]. The approach is not strictly a CTS one since it does not require any complex annotation schema or strong coupling between NLG and TTS. Instead, the semantic representation of the dialogue acts is used to extract context features for decision tree state clustering. The work reported here continues in this direction by demonstrating that dialogue context sensitive speech synthesis can be applied to different domains without needing to record domain specific training/adaptation data.

3. Context Adaptive Training with Factorized Decision Trees

The realization of an utterance is governed by a number of strong contextual factors such as the phonetic content and coarticulations (which strongly affect the spectrum and also influence prosody) and the stress, accent and tone allocation (which have a strong effect on the fundamental frequency and can also impact spectral content). However, the realization of an utterance is also affected by weak contextual factors such as style and emphasis. The effects of all the contextual factors are coupled at both the segmental and suprasegmental levels. All such factors must be considered for HMM-based speech synthesis in order to achieve a high quality output. However, when a decision tree is used to cluster and model similar realizations, the influence of each factor may not be modelled accurately due to sparsity of the available data for some of the contextual factors.

Context Adaptive Training with Factorized Decision trees (FD-CAT) was introduced to jointly model the influence of both strong and weak context factors [23]. The contextual factors are

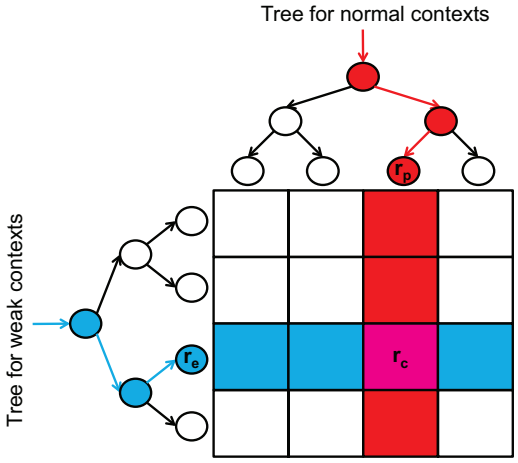

Figure 1: *Factorized Decision Trees.*

divided into two sets, one comprising the more influential factors (usually the normal set of HTS contexts) while the other contains the weak factors that contribute less to the likelihood of the data. Full context-dependent HMMs are trained using the normal set of contexts, and then adapted using the weak-context specific transformations. Two separate decision trees are trained on the same data set using the normal and weak context factors. These are then combined into a single decision tree with intersected leaf nodes as shown in Fig. 1.

Following the notation in [10], each leaf node $r_p = \theta_1, \cdots, \theta_{N_p}$ from the normal-context decision tree is combined with each leaf node $r_e = \theta_1, \cdots, \theta_{N_e}$ from the weak-context decision tree into a leaf node r_c in the combined decision tree

$$r_c = r_p \cap r_e := \{\theta : \theta \in r_p \wedge \theta \in r_e\} \quad (1)$$

where θ is a distinct state corresponding to a full-context model. The leaf nodes of the combined tree r_c form atomic adaptation units on which both r_p and r_e will have an effect. For the work here, an implementation based on MLLR adaptive training was used whereby the mean and covariance of the Gaussian component m are given by

$$\hat{\mu}_m = A_{r_e(m)}\mu_{r_p(m)} + b_{r_e(m)}\,, \quad \hat{\Sigma}_m = \Sigma_{r_p(m)} \quad (2)$$

where $r_p(m)$ and $r_e(m)$ are the leaf nodes of the normal and weak context decision trees containing m, $A_{r_e(m)}$ and $b_{r_e(m)}$ are the weak context transform parameters and $\mu_{r_p(m)}$ and $\Sigma_{r_p(m)}$ are the normal context Gaussian parameters. More details about the implementation of FD-CAT can be found in [10].

4. Data Description

The speech corpus used in this study consists of a general purpose text-to-speech corpus combined with an expressive dialogue corpus specifically designed for the TownInfo domain. The former was provided by Phonetic Arts for research purposes [27] and is referred to in that corpus as the RJS voice. The latter was recorded by the same RJS speaker using prompt scripts derived from the logs of the Cambridge spoken dialogue system in the TownInfo and TopTable restaurant domains [9]. The TownInfo domain includes restaurant, hotel and bar information for a hand-crafted database, while the TopTable domain contains restaurants provided by an online service provider [28].

Emphasis and style were selected as the primary expressive patterns to be covered. In order to provide emphasis data

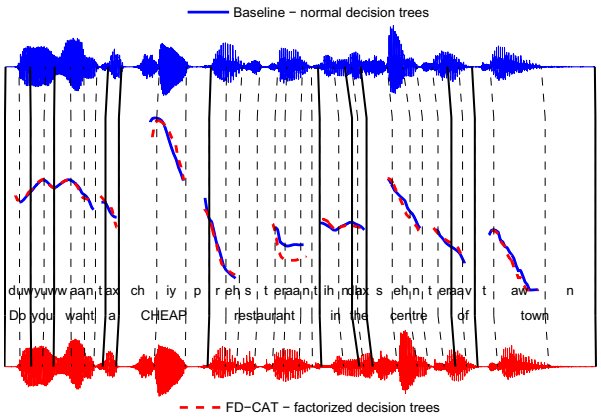

Figure 2: *F0 contours of a restaurant-domain utterance. The blue solid contour corresponds to the sentence synthesized with the baseline voice while the red dashed contour was produced by the FD-CAT voice. Both contours were time-warped in order to align the phoneme boundaries (the mean values were used as the reference points). The original waveforms are also plotted above for the baseline and below for the FD-CAT sentences. Word boundaries and phoneme boundaries are shown with vertical solid and dashed lines respectively.*

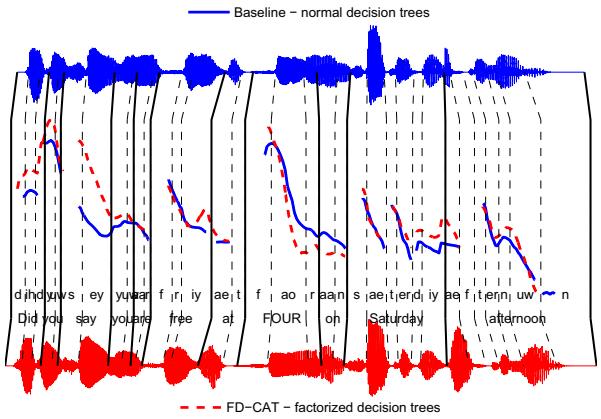

Figure 3: *F0 contours of a booking-domain utterance synthesized with the baseline (solid) and the FD-CAT (dashed) voices.*

the dialogue corpus was annotated using emphasis tags at the slot level using the following simple approach. For every dialogue the first encounter of each slot value was annotated with an emphasis tag. Expressive style, on the other hand, is neither precisely defined, nor is an annotation scheme available for it. Hence, the style was modeled implicitly by including whole dialogues in the corpus. The speaker was instructed to take on the role of the system for each given dialogue task, as well as to follow the emphasis annotations as closely as possible. The collected expressive dialogue corpus contains 3158 wave files totalling about 5 hours of audio. Each wave file corresponds to a system-uttered prompt consisting of one or more sentences and is associated with the emphasis-tagged text prompt, as well as the dialogue act that was used to generate it.

Due to repetition of slot values and limited vocabulary, the resulting corpus is not phonetically balanced. Consequently, the use of a simple decision tree for this data leads to unbalanced modeling of the contextual factors, e.g. the phoneme sequence of a slot value that appears several times emphasized (or in a specific dialogue context) can be modeled adequately, while other sequences of phonemes may lack emphatic (or dialogue) context and cannot be accurately modeled.

5. Experiments

5.1. Domain description

The in-domain task involves restaurant information system and the out-of-domain task involves appointment bookings. The ontologies, i.e. the entities, slots and values, are widely different for the selected tasks. However, both domains share the same scheme for dialogue act specification. Dialogue acts take the form $dact(a_1[=v_1], \ldots, a_N[=v_N])$, where $dact$ is the dialogue act type, $\{a_i, v_i\}$ is the i-th slot-value pair, and N is the number of slots, e.g. *confirm(pricerange=cheap,area=centre)* realized as *"Do you want a cheap restaurant in the centre of town?"*, or *confirm(date=Saturday,hour=four,period=pm)* from the booking domain realized as *"Did you say you are free at four on Saturday afternoon?"*. Minor differences exist in the set of act types, e.g. *confbook* - confirming the appointment date and time before asking whether to book it. Such act types have similar counterparts in the training data and were mapped accordingly.

5.2. Dialogue Context Sensitive Voices

As described above, two dialogue context sensitive voices were built: a baseline using standard decision trees and one using FD-CAT. The baseline uses standard decision trees with the emphasis and dialogue context features combined with the standard set of linguistic features. The FD-CAT system splits out the emphasis and dialogue features as a separate set of weak factors. The weak context question set had a total of 23 questions, six of which were derived from the emphasis context and the rest from the dialogue act semantic representation [9]. Both the baseline and the FD-CAT voices were trained on the same data including the original RJS corpus and the dialogue corpus. The following stream configuration was used: 25 Mel-Cepstral coefficients, five band aperiodic energy components, continuous log F0 and voicing condition [29]. A modified version of the HTS framework was used for FD-CAT as described in [10].

Figures 2 and 3 illustrate the differences between the two voices for in-domain and out-of-domain utterances, respectively. Each figure shows a sentence from the corresponding domain synthesized with the baseline (top) and the FD-CAT (bottom) voices. The F0 contours are shown after being aligned at the phoneme boundaries which are marked with vertical lines. For the in-domain case (Fig. 2), there are few observed differences with the duration patterns being very similar and the F0 contours coincide most of the time. On the other hand, there are significant differences in both the duration patterns and the F0 contours for the out-of-domain case (Fig. 3), e.g. the duration of the emphasized word FOUR is perceivably longer for the FD-CAT voice, and the F0 contour is higher for the initial words of the utterance. This suggests that the FD-CAT approach has a greater effect for the out-of-domain task while the effect for the in-domain case is limited since the required realizations are already well represented in the training data. This was also observed in the listening test presented in the following section.

5.3. Evaluation

A preference listening test designed for dialogue appropriateness was performed to evaluate the baseline and FD-CAT voices for the in-domain and out-of-domain tasks. The listener was presented with a dialogue script including both the system

prompts and the user responses. The top ASR hypothesis was used as the user response instead of the actual user's speech transcription so that the listener is not affected by any misrecognitions. The emphasized words were marked in bold-face and listeners were instructed to take this into account in order to factor out the effect of the emphasis assignment algorithm. The listener was asked to choose the most appropriate between the two alternative synthesized versions for each turn or indicate "no preference". The presentation order was randomized, and the listeners were allowed to playback the audio multiple times.

A set of 20 previously collected dialogues were randomly selected for the restaurant domain. The system used to collect them integrated an emphasis assignment module so the prompts already contained emphasis annotations. For the appointment booking domain a set of 15 dialogues were selected from an existing corpus and then annotated so that the first occurrence of each slot value was emphasized. Each task was evaluated three times using subjects recruited via the Amazon Mechanical Turk crowd-sourcing service constrained to ensure only one task per listener per domain. The listeners were also asked to report if they were native or non-native English speakers.

The results are summarized in Table 1; the upper section refers to the restaurant domain and the lower section to the appointment booking domain. The table also shows the breakdown according to whether the prompt contained an emphasized slot (*emphasis*) or not (*plain*), as well as whether the listener reported being *native* or *non-native*. A breakdown according to the dialogue act is also shown grouping them in three categories (*confirm*: the system confirms a slot, *inform*: informs one or more slots, and *request*: requests information). For each comparison the number of judgements is shown as well as the statistical significance level computed using a one-sided binomial test after equally splitting the "no preference" values. The statistical tests are clearly not independent from one another, however the number of samples does not favour testing each combination of the three conditions separately.

As can be seen, there is no significant preference towards either the baseline or the FD-CAT voice for the in-domain task. This holds for both the plain and emphasis scenario, native and non-native speakers as well as across dialogue act types. This is in accordance with the observations made in Sec. 5.2 and verifies the presumption that adaptive training has little effect for in-domain utterances since there is sufficient coverage in the training data. In contrast, the FD-CAT voice is consistently preferred over the baseline for the out-of-domain task. There is good agreement between the native and non-native speakers' judgements, though the difference for the latter is not statistically significant due to the lack of samples. The preference is statistically significant for the plain scenario. This is partly due to the request prompts which are usually questions without emphasized slots. The preference in the emphasis scenario as well as the inform and confirm cases is not statistically significant. The combined effect of emphasis and dialogue contexts does not manifest itself. This suggests that the data coverage did not affect significantly the emphasis generalization for the out-of-domain task while it is more important the dialogue context.

The FD-CAT voice was also compared to a voice with neutral characteristics which was trained on the RJS corpus without including the expressive dialogue data. The results summarized in Table 2 show that the FD-CAT voice is significantly preferred over the neutral voice. This demonstrates that context-sensitive speech synthesis is more appropriate for dialogue than a voice with neutral characteristics, and can be applied to different domains without requiring additional training/adaptation data.

Restaurant Information Task (in-domain)					
Condition	# Judg.	Baseline	No Pref.	FD-CAT	p-value
plain	198	29.3%	43.4%	27.3%	0.415
emphasis	201	39.3%	22.9%	37.8%	0.443
non-native	231	36.8%	27.7%	35.5%	0.447
native	168	31.0%	40.5%	28.6%	0.408
inform	258	34.1%	32.2%	33.7%	0.475
confirm	93	37.6%	25.8%	36.6%	0.500
request	48	29.2%	52.1%	18.8%	0.235
Total	399	34.3%	33.1%	32.6%	0.381
Appointment Booking Task (out-of-domain)					
Condition	# Judg.	Baseline	No Pref.	FD-CAT	p-value
plain	162	27.8%	24.7%	**47.5%**	**0.007**
emphasis	195	38.5%	15.9%	45.6%	0.158
non-native	132	32.6%	21.2%	46.2%	0.069
native	225	34.2%	19.1%	**46.7%**	**0.030**
inform	222	37.8%	18.5%	43.7%	0.191
confirm	63	31.7%	25.4%	42.9%	0.224
request	72	22.2%	19.4%	**58.3%**	**0.001**
Total	357	33.6%	19.9%	**46.5%**	**0.007**

Table 1: Preference results comparing the baseline to the FD-CAT voice. Significant results are shown in bold ($p<0.05$).

Domain	# Judg.	Neutral	No Pref.	FD-CAT	p-value
Restaurant	399	32.1%	9.3%	**58.6%**	**4.7E-08**
Booking	357	34.5%	10.6%	**54.9%**	**6.6E-05**

Table 2: Preference results comparing the FD-CAT dialogue context sensitive voice to a voice with neutral characteristics.

6. Conclusions

The quality of the synthesized speech from a Spoken Dialogue Systems is critical to both intelligibility and naturalness. In addition to accurate articulation, the prosody and style must be both realistic and appropriate to the changing dialogue context. Recently, we introduced dialogue context features for decision tree state clustering and observed significant improvements in the perceived quality and contextual appropriateness of the synthetic voice [9]. This approach relies on the use of training data which includes all of the required emphases and styles. The HTS decision trees are then trained to map the explicit context features, in this case dialogue acts and emphasized words, onto the corresponding patterns in the training data.

An obvious limitation of this approach is the reliance on in-domain training data. When used out-of-domain, the training data is inevitably phonetically unbalanced and simple decision trees cannot capture the expressive patterns in the data. Hence, the work reported in this paper has sought to determine how significant this problem is, and investigate the extent to which the separation of strong and weak features using the FD-CAT HMM synthesis approach can mitigate the problem.

Experiments have been conducted to compare a standard voice trained using a simple decision tree to a voice trained using FD-CAT. Both voices were tested on an in-domain Restaurant information task and an out-of-domain appointments booking task. The results show that both voices are equally preferable for the in-domain task, whereas for the out-of-domain task the FD-CAT voice is significantly preferred over the baseline.

7. References

[1] S. Young, "Still talking to machines (cognitively speaking)," in *Proceedings of INTERSPEECH*, 2010, pp. 1–10.

[2] H. Hastie et al, "Demonstration of the Parlance system: a data-driven, incremental, spoken dialogue system for interactive search," in *SIGDIAL 2013*.

[3] G. Aist, J. Allen, E. Campana, and C. G. Gallo, "Incremental understanding in human-computer dialogue and experimental evidence for advantages over nonincremental methods," in *Proceedings Workshop on the Semantics and Pragmatics of Dialogue (DECALOG)*, 2007.

[4] D. Schlangen and G. Skantze, "A general, abstract model of incremental dialogue processing," in *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2009, pp. 710–718.

[5] E. O. Selfridge, P. A. Heeman, I. Arizmendi, and J. D. Williams, "Demonstrating the incremental interaction manager in an end-to-end "lets go!" dialogue system," in *Proc. of IEEE Workshop on Spoken Language Technology*, 2012.

[6] N. Dethlefs, H. Hastie, V. Rieser, and O. Lemon, "Optimising incremental generation for spoken dialogue systems: reducing the need for fillers," in *Proceedings of the Seventh International Natural Language Generation Conference*. Association for Computational Linguistics, 2012, pp. 49–58.

[7] T. Baumann and D. Schlangen, "Evaluating prosodic processing for incremental speech synthesis," in *Proceedings of INTERSPEECH*, 2012.

[8] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, "Speech synthesis based on hidden markov models," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1234–1252, May 2013.

[9] P. Tsiakoulis, C. Breslin, M. Gašić, M. Henderson, D. Kim, M. Szummer, B. Thomson, and S. Young, "Dialogue context sensitive HMM-based speech synthesis," in *Proceedings of ICASSP*, 2014.

[10] K. Yu, H. Zen, F. Mairesse, and S. Young, "Context adaptive training with factorized decision trees for HMM-based statistical parametric speech synthesis," *Speech Communication*, vol. 53, no. 6, pp. 914–923, 2011.

[11] S. Young and F. Fallside, "Speech synthesis from concept: a method for speech output from information systems," *The Journal of the Acoustical Society of America*, vol. 66, p. 685, 1979.

[12] J. Hitzeman, A. W. Black, P. Taylor, C. Mellish, and J. Oberlander, "On the use of automatically generated discourse-level information in a concept-to-speech synthesis system," in *Proceedings of ICSLP*, 1998.

[13] S. Takada, Y. Yagi, K. Hirose, and N. Minematsu, "A framework of reply speech generation for concept-to-speech conversion in spoken dialogue systems," in *Proceedings of INTERSPEECH*, 2007, pp. 1286–1289.

[14] P. A. Taylor, "Concept-to-speech synthesis by phonological structure matching," *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, vol. 358, no. 1769, pp. 1403–1417, 2000.

[15] I. Bulyko and M. Ostendorf, "Efficient integrated response generation from multiple targets using weighted finite state transducers," *Computer Speech & Language*, vol. 16, no. 3, pp. 533–550, 2002.

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

[17] S. Pan, "Prosody modeling in concept-to-speech generation," Ph.D. dissertation, Columbia University, 2002.

[18] M. Schnell and R. Hoffmann, "What concept-to-speech can gain for prosody," in *Proceedings of INTERSPEECH*, 2004.

[19] J. Yamagishi, T. Masuko, and T. Kobayashi, "A style control technique for HMM-based expressive speech synthesis," *IEICE TRANSACTIONS on Information and Systems*, vol. 90, no. 9, pp. 1406–1413, 2007.

[20] L. Chen, M. J. Gales, V. Wan, J. Latorre, and M. Akamine, "Exploring rich expressive information from audiobook data using cluster adaptive training," in *Proceedings of INTERSPEECH*, 2012.

[21] H. Zen, N. Braunschweiler, S. Buchholz, M. J. Gales, K. Knill, S. Krstulovic, and J. Latorre, "Statistical parametric speech synthesis based on speaker and language factorization," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 6, pp. 1713–1724, 2012.

[22] L. Badino, J. S. Andersson, J. Yamagishi, and R. A. Clark, "Identification of contrast and its emphatic realization in HMM-based speech synthesis," in *Proceedings of INTERSPEECH*, 2009.

[23] K. Yu, F. Mairesse, and S. Young, "Word-level emphasis modelling in HMM-based speech synthesis," in *Proceedings of ICASSP*. IEEE, 2010, pp. 4238–4241.

[24] K. Morizane, K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Emphasized speech synthesis based on hidden markov models," in *Proceedings of Oriental COCOSDA International Conference on Speech Database and Assessments*. IEEE, 2009, pp. 76–81.

[25] F. Meng, Z. Wu, J. Jia, H. Meng, and L. Cai, "Synthesizing english emphatic speech for multimodal corrective feedback in computer-aided pronunciation training," *Multimedia Tools and Applications*, pp. 1–27, 2013.

[26] Y. Maeno, T. Nose, T. Kobayashi, T. Koriyama, Y. Ijima, H. Nakajima, H. Mizuno, and O. Yoshioka, "HMM-based expressive speech synthesis based on phrase-level F0 context labeling," in *Proceedings of ICASSP*, 2013, pp. 7859–7863.

[27] S. King and V. Karaiskos, "The Blizzard Challenge 2010," 2010.

[28] P. Tsiakoulis, M. Gašić, M. Henderson, J. Planells-Lerma, J. Prombonas, B. Thomson, K. Yu, S. Young, and E. Tzirkel, "Statistical methods for building robust spoken dialogue systems in an automobile," in *4th Int'l Conf. on Applied Human Factors and Ergonomics*, 2012.

[29] K. Yu and S. Young, "Continuous F0 modeling for HMM based statistical parametric speech synthesis," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 5, pp. 1071–1079, 2011.