DIALOGUE CONTEXT SENSITIVE HMM-BASED SPEECH SYNTHESIS

Pirros Tsiakoulis, Catherine Breslin, Milica Gašić, Matthew Henderson, Dongho Kim, Martin Szummer, Blaise Thomson, Steve Young

University of Cambridge, Engineering Department, Cambridge, UK

pt344@cam.ac.uk

ABSTRACT

The focus of this work is speech synthesis tailored to the needs of spoken dialogue systems. More specifically, the framework of HMM-based speech synthesis is utilized to train an emphatic synthetic voice that also considers dialogue context for decision tree state clustering. To achieve this, we designed and recorded a speech corpus comprising of system turns from human-computer interaction, as well as additional prompts for slot-level emphasis. This corpus, combined with a general purpose text-to-speech one, was used to train HMM-based synthetic voices using a) baseline context features, b) additional slot-level emphasis features, and c) additional dialogue context features extracted from the dialogue act semantic representation. The voices were evaluated in pairs for dialogue appropriateness using a preference listening test. The results show that the emphatic voice is more preferable than the baseline when emphasis markup is present, while the dialogue context-sensitive voice is more preferable than the plain emphatic one when no emphasis markup is present and more preferable than the baseline in both cases.

Index Terms: HMM-based speech synthesis, emphatic speech synthesis, dialogue context, context sensitive speech

1. INTRODUCTION

Speech has gained significant ground as a human-machine interface, enabling Spoken Dialogue Systems (SDS) for a variety of applications [1]. Such systems often employ a general purpose synthetic voice with neutral characteristics. Recent effort has focused on making the discourse more natural, incorporating spontaneous responses, backchannel and fillers, as well as incremental processing [2, 3, 4, 5, 6, 7]. This pinpoints the need for expressive speech synthesis that is aware of the discourse context [8]. The generated system prompts need to be concise and convey more information via prosody. The Text-to-Speech (TTS) component of a spoken dialogue system is typically preceded by the Natural Language Generation (NLG) component. The NLG component translates the intended dialogue action from a high-level semantic representation into text. This facilitates richer generation; in addition to plain text output, the NLG component can also produce expressive annotations [9, 10]. However, expert knowledge and effort is required to design and implement both the NLG and TTS components.

This paper investigates the potential of an expressive TTS component targeting the needs of a spoken dialogue system without the need of any complex annotation scheme. Instead, the existing dialogue act semantic representation is used as an additional contextual factor for decision tree state clustering in HMM-based speech synthesis. This work mainly considers emphasis and style as the target aspects of expressive speech for dialogue. Emphasis provides a way of highlighting the focus of the utterance and naturally signalling what the user should pay attention to. Style, on the other hand, which can be manifested in various ways, e.g. speaking rate, pitch variations, etc., can be used to convey more subtle information to the user. For example, the speaking rate may be reduced (in conjunction with emphasis) when giving new information to the user. Both emphasis and style context features are generated from the input dialogue act.

To this end, a new speech corpus was collected to be used for expressive speech generation within the dialogue domain. The corpus includes *system-user* pairs of interaction prompts from previously collected dialogues, as well as individual prompts designed specifically for emphasis patterns. A professional speaker was instructed to act as the dialogue system operator and convey information to the *user* using contextually appropriate speech. The collected speech corpus was used in addition to a general purpose text-to-speech corpus to build: a) a voice using baseline context features, b) an emphatic voice by including slot-level emphasis context features, and c) a dialogue context-sensitive emphatic voice by including contextual factors extracted from the intended dialogue act semantic representation. A special preference listening test was designed to evaluate the voices in the context of a spoken dialogue system. A dialogue was presented to the user where each system turn had a pair of alternative synthetic prompts. The user was asked to choose the most appropriate system response or a third choice if he or she had no preference. The results show that the emphatic voice is preferred to the baseline, when emphasis markup is present, while the context-sensitive voice is preferred to the plain emphatic one, when no emphasis markup is present, and preferred to the baseline in both cases.

1.1. Related Work

The idea of semantic input to the speech synthesizer was originally introduced by Young and Fallside using the term *Speech Synthesis from Concept* [11]. The term *Concept-To-Speech* later prevailed to describe methods that combine joint NLG and TTS functionality. One approach to CTS involves an annotation schema which is applied to the generated text, and affects the prosody of the rendered speech [9]. A similar technique applies prosodic annotations to a template-slot based generation system [10]. Another approach is to jointly optimize text and prosody generation in the framework of unit selection concatenative speech synthesis [12, 13]. Others have focused on prosody models for CTS, which are driven from semantic input as well as linguistic input [14, 15, 16]. Our 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 in HMM-based speech synthesis.

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

There is a considerable amount of ongoing research on HMM based statistical speech synthesis (HTS) [17], which has led to significant improvement in the quality of the synthetic speech [18]. 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. In this paper, the HTS framework is utilized to investigate the use of dialogue and emphasis information that is directly extracted from the dialogue act representation.

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 [19, 20] or by collecting speech corpora specifically designed for emphasis modeling [21]. Emphasis context features are then used in the decision tree state clustering stage. More elaborate techniques have also been proposed that can tackle data sparsity issues when the emphasis data is limited, such as factorized decision trees [20, 22], hierarchical modeling [23], and phrase level modeling [24]. Adaptation techniques have also been proposed for different aspects of expressive speech synthesis [25, 26]. The goal of this work is not to propose a new technique, but rather explore existing ones in the context of a dialogue system.

2. EXPRESSIVE DIALOGUE CORPUS

The restaurant domain was selected as the primary application domain, mainly because of data availability. Emphasis and style were selected as the primary expressive patters to be covered. To achieve this, the scripts to be recorded were annotated to indicate those words that should be emphasized. The expressive style, on the other hand, is neither strictly defined, nor an annotation format is available. Specifically for the dialogue domain, the expressive style should reflect the current dialogue state, e.g. the confidence level. These phenomena were modeled implicitly by including whole dialogues into the corpus.

2.1. Existing Dialogue Corpora

The initial source data consisted of previously collected dialogues using the Cambridge spoken dialogue system. This data is summarized in Table 1. The TownInfo domain includes restaurant, hotel and bar information for a hand-crafted information database [27], while the TopTable domain is for restaurants provided by an online service provider [28]. An initial investigation into using a subset of this dialogue data directly showed that it was not rich enough for the purpose at hand. It had very limited diversity in terms of the system prompts as well as the available venue names. Note that the counts the unique prompts in Table 1 were calculated including the actual slot values used. Therefore it was decided to preprocess the prompts and enrich them.

Domain	# Dialogues	# Turns	# Unique Prompts
TownInfo	1422	13992	3346
TopTable	2166	28846	2284
Total	3588	42838	5614

Table 1. Source dialogue data for the corpus design.

Domain	# Dialogues	# Turns	# Unique Prompts
TownInfo	86	1089	407
TopTable	131	1351	1018
Total	217	2440	1425

Table 2. Dialogue data included in the final corpus.

2.2. Prompt Processing and Corpus Selection

In order to add some variety into the desing of the final corpus, the extracted prompts were enriched semi-automatically. The transformation procedure included the following steps:

- Extract pairs of dialogue act and corresponding system prompt, e.g.
  `inform(name="st johns chop house", postcode="CB3 0AD")`
  `The postcode of st johns chop house is cb3 0ad`
- Replace slot values with slot class names, e.g.
  `inform(name=NAME, postcode=POSTCODE)`
  `The postcode of NAME is POSTCODE`
- Manually check the unique list of the templates and provide alternatives by rephrasing the prompt, e.g.
  `inform(name=NAME, postcode=POSTCODE)`
  `The postcode of NAME is POSTCODE`
  `Its postcode is POSTCODE`
  `The postal code of NAME is POSTCODE`
  `...`
- Select a list of dialogues that maximize the coverage of the extended list of prompts. A simple greedy algorithm was used for this task. At each step, the algorithm added to the list the dialogue which included the most unseen prompts. This is similar to the standard approaches to corpus design that operate at the word level [29].
- The slot class names were replaced with slot values. If a dialogue involved a venue that had already been spoken about in a previously selected dialogue, the venue was randomly replaced with another venue to avoid many repetitions of the same venue name. For other slots, such as phone number and postcode, a random list was generated.
- Some artificial turns were added to include free text descriptions that were available for some of the venues.

The summary of the selected dialogues is shown in Table 2.

2.3. Emphasis Assignment

Prosodic emphasis is an important expressive pattern for the speech synthesis component of the dialogue system. The dialogue corpus was annotated using emphasis tags at the slot level. More specifically, for every dialogue the first encounter of each slot value was marked and annotated with an emphasis tag. This assumes that the system should emphasize every new bit of information that it presents to the user[1]. The technique was also implemented and integrated into the dialogue system which was used for evaluation. An example dialogue is shown below where emphasized words are marked in bold-face.

[1]More elaborate methods can be applied at run-time by the NLG component to assign slot/word emphasis tags.

Prompt:	Hello, how may I be of assistance?
User:	*I'm looking for a moderately priced restaurant in the eastern part of town.*
Prompt:	There are **41 restaurants** in the **moderate** price range. Do you have any preference regarding the cuisine?
User:	*I don't care, any that's fine.*
Prompt:	Alright, the type of food doesn't matter. Which area do you have in mind?
User:	*In the east of town please.*
Prompt:	**Rajmahal** is a quiet restaurant in the **east** of Cambridge in the moderate price range.
User:	*What's the address and type of food of the restaurant?*
Prompt:	Sure, Rajmahal is on **7 Barnwell Road Fen Ditton**.
User:	*And what type of food does it serve?*
Prompt:	Rajmahal serves **Indian**.
User:	*Can you tell me some more about it?*
Prompt:	Rajmahal serves traditional Indian cuisine accompanied by a broad wine list.
User:	*OK, thank you very much, goodbye*
Prompt:	I hope I was helpful.

In addition, an extra set of prompts was selected for the emphasis task. Each prompt was recorded multiple times, each time with a different slot emphasized. An example is given below

Are you looking for a Portuguese restaurant in Barnwell ?
Are you looking for a **Portuguese** restaurant in Barnwell ?
Are you looking for a Portuguese **restaurant** in Barnwell ?
Are you looking for a Portuguese restaurant in **Barnwell** ?

2.4. Speech Recordings

The RJS speech corpus was used as a starting dataset to train HMM-based synthetic voices [30]. In order to utilize this data, the expressive dialogue corpus was recorded by the same speaker. The recording took place in a professional studio in four three-hour sessions during a period of four days. The speaker was instructed to take on the role of the operator for each given dialogue task, as well as to follow the emphasis annotations as closely as possible.

The raw audio data was then split into smaller wave files, each containing the prompt of a single system turn. Each turn may contain more than one sentence. The reason for this choice is to model the prosody of each turn as a whole, utilizing dialogue context features. Each wave file is associated with the emphasis-tagged text prompt, as well as the dialogue act that was used to generate it. The final corpus contains 3158 wave files totalling about 5 hours of audio. A set of 50 sentences from the existing recordings were also recorded. The difference between the average spectrum of the original 50 sentences compared to that of the corresponding new recordings was used as a spectrum equalizer for the new dataset.

3. EXPERIMENTS

3.1. Voice configurations

All voices were trained on the same dataset including the original RJS corpus and the new expressive dialogue one. The training setup was also kept the same using a modified version of the HTS framework that incorporates continuous F0 contour modeling [31]. The following stream configuration was used: 25 Mel-Cepstral coefficients, log F0, five frequency band aperiodic energy components, and voicing condition [31]. The STRAIGHT vocoder was used for speech analysis and feature extraction.

Three voices were trained: a) a voice using baseline context features for state clustering, b) an emphatic voice by including emphasis context features (6 additional context questions were used [20]), and c) a dialogue context-sensitive emphatic voice by including three dialogue act features; the dialogue act type (17 additional questions), the number of slots in the dialogue act (11 additional questions), and one additional question whether the act was a negative inform or not.

3.2. Live experiment

A live experiment was carried out using crowd-sourcing via Amazon Mechanical Turk. The users were asked to call a toll-free number and talk to the dialogue system. For each call, the user was assigned a randomly generated task from the TopTable domain, one or more slots were specified, e.g. food type, price range, area, etc. The user had to negotiate with the system to get a venue matching the given constraints, and had to ask and get specific information about that venue, e.g. address, phone number, cuisine, etc. More complex dialogues would occur if there were no matching venues, in which case the user could relax one of the given constraints. At the end of the dialogue, the user was asked to judge the dialogue for: a) task completion success (*Yes* or *No*), b) perceived comprehension on a five-point Likert scale from *strongly disagree* to *strongly agree* (if the system understood the user), c) overall impression for the quality of the systems voice, d) emphasis assignment, and f) intonation. The later three questions, which are relevant for TTS, were rated on a continuous scale (0-60) for Mean Opinion Score (MOS) [18].

Three systems were tested having identical configurations, except the synthetic voice used by the TTS component. Each user could make up to 15 calls, and each call was randomly routed to one of the available systems. A total of 274 dialogues were collected from 26 users after discarding those who did not speak to all three systems. The results are shown in Table 3. None of the differences is statistically significant. Moreover, the MOS responses (Overall, Emphasis, and Intonation) are highly correlated to each other ($>$0.8) and moderately correlated to Comprehension (0.40, 0.29, 0.33).

Voice	Succ.	Compr.	Overall	Emphasis	Intonation
`base`	91.9%	3.65	42.6	41.3	41.8
`emph`	90.3%	3.85	42.2	40.6	40.5
`dact`	89.0%	3.78	42.4	41.7	41.7

Table 3. Live experiment results. Each row corresponds to a synthetic voice and each column to the question asked.

Analysis of Variance (ANOVA) was performed on the data, in order to discover which factors affected the users' responses. Table 4 summarizes the one-way ANOVA results for each of the MOS answers against different factors[2]. The results show that Comprehension is the most significant factor in explaining the variance for all the MOS observations. Success, having moderate correlation with Comprehension (0.38), is also a significant factor. On the other hand, the Voice factor has no significant effect on any of the Overall, Emphasis, or Intonation MOS factors. The results show that the desig of the experiment was not effective in assessing the utility of the synthetic voices. The users could not disentangle the primary task of maintaining the dialogue to find a venue from the secondary task of evaluating the quality of the synthetic speech.

[2]Two-way and three-way ANOVA was also performed, however no significant effect was found for any combination involving Voice factor.

Factor	Overall	Emphasis	Intonation
Voice	0.029 (0.971)	0.161 (0.851)	0.222 (0.801)
Succ.	5.456 (0.020)	1.079 (0.300)	3.841 (0.051)
Compr.	47.47 (<0.001)	22.29 (<0.001)	31.25 (<0.001)

Table 4. One-way ANOVA results of the MOS answers compared to the Voice, Success, and Comprehension factors. The F-value is shown for each pair as well as, the significance level (p-value).

3.3. Listening test

Given the above, a special preference listening test was designed to evaluate the three voice setups in the context of a spoken dialogue system. The listerer was presented a dialogue 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. Each system turn had a pair of alternative synthetic prompts, and the listener was asked to choose the most appropriate one or indicate no preference. One could listen to each pair mutliple times, though this happens rarely with crowd-sourced evaluators.

The voices were evaluated in pairs. A set of 50 dialogues were randomly selected from the ones collected during the live listening test. The system prompts were synthesized with all the three voice setups, using the actual dialogue acts and the emphasis tags that were assigned at runtime. For each dialogue three listening tasks were generated (one per pair). The presentation order of the two synthetic prompts was randomized within each task. Each task was evaluated at most 6 times via crowd-sourcing. A total of 339 evaluators completed the listening tasks, resulting in a total of 6395 judgements.

The results are summarized in Table 5 organized in three sections. The top section compares the baseline voice versus the emphatic voice, the mid section compares the emphatic voice to the dialogue context-sensitive voice, and the last section compares the baseline to the context-sensitive voice. For each comparison, the total preference percentages are shown, as well as the breakdown according to two conditions. The first one is whether the prompt contained an emphasized slot (*emphasis*) or not (*plain*), while the other breaks down the results according to the dialogue act type (*confirm* - the system is confirming a slot, *confreq* - confirming a slot while requesting another, *inform* - informing one or more slots, and *request* - requesting information for a slot). The number of judgements per comparison is also shown, as well as the statistical significance level estimated using a sign test.

The comparison between the baseline and the emphatic voice shows significant preference towards the emphatic one. This preference is mainly attributed to the sentences containing emphasized slots, while there is insignificant preference to the baseline voice in case of prompts without emphasis (*plain*). The preference is also significant for the inform dialogue act. This is expected since more than half of the total number of prompts were of inform type and about half of them contained emphasized slots. The comparison between the emphatic voice and the context-sensitive one shows significant preference towards the latter, when there is no emphasis present, while there is no preference otherwise. Moreover, the latter is more preferable for all the different the dialogue act types (significantly for the confirm and request types). The final comparison shows significant preference for the context-sensitive voice compared to the baseline regardless of the emphasis presence. The preference is significant for the inform and confreq dialogue acts.

The results largely agree with the intuition given the training setup. The emphasis factor makes a difference only for emphasized prompts (both emphatic and context-sensitive voices are significantly prefered to the baseline), otherwise there is no effect (no significant difference between baseline and emphatic in the plain scenario, or between emphatic and context-sensitive in the emphasis scenario). The dialogue context factor, on the other hand, has a positive effect regardless of the emphasis status while maintaining the emphasis advantage, since it improves on both the baseline and the emphatic voice for most of the dilaogue act types.

Baseline versus Emphatic					
Condition	# Judg.	Baseline	Neutral	Emphatic	p-value
plain	1138	34.3%	34.4%	31.4%	0.164
emphasis	1012	32.0%	23.4%	**44.6%**	<0.001
confirm	180	36.7%	31.1%	32.2%	0.301
confreq	317	37.9%	22.1%	40.1%	0.368
inform	1338	33.6%	25.6%	**40.8%**	0.005
request	315	24.8%	50.8%	24.4%	0.500
Total	2150	33.2%	29.2%	**37.6%**	0.022
Emphatic vs Dialogue					
Condition	# Judg.	Emphatic	Neutral	Dialogue	p-value
plain	1116	28.8%	33.2%	**38.0%**	0.001
emphasis	1008	38.8%	21.9%	39.3%	0.437
confirm	180	33.3%	17.2%	**49.4%**	0.015
confreq	312	38.8%	19.6%	41.7%	0.305
inform	1320	35.5%	26.1%	38.4%	0.154
request	312	19.9%	50.0%	**30.1%**	0.039
Total	2124	33.5%	27.9%	**38.6%**	0.010
Baseline vs Dialogue					
Condition	# Judg.	Baseline	Neutral	Dialogue	p-value
plain	1115	23.9%	46.7%	**29.3%**	0.036
emphasis	1006	27.1%	28.8%	**44.0%**	<0.001
confirm	180	30.0%	33.9%	36.1%	0.206
confreq	312	21.8%	32.1%	**46.2%**	<0.001
inform	1317	27.0%	35.7%	**37.4%**	<0.001
request	312	20.2%	57.7%	22.1%	0.388
Total	2121	25.5%	38.2%	**36.3%**	<0.001

Table 5. Preference results comparing the three synthetic voices in pairs. Significant results are shown in bold ($p<0.05$).

4. CONCLUSIONS

An expressive dialogue corpus which contains in-domain examples of context-sensitive prosody spoken by a professional speaker has been designed and collected. Based on this corpus, a prototype has been developed that generates context-sensitive emphasis and prosody. The prototype incorporates a simple algorithm that emphasizes every new bit of information that presents to the user, as well as an HMM-based synthetic voice that was trained with both emphasis and dialogue context features for decision tree state clustering. This prototype voice was evaluated in contrast to two alternatives, i.e. one that was trained with baseline context features, and another that additionally incorporated emphasis context features. The results show that there is significant preference for the context-sensitive voice in a listening test for dialogue. Future work will investigate the combination of additional dialogue context features with more advanced training techniques.

5. REFERENCES

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

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

[3] Gregory Aist, James Allen, Ellen Campana, and Carlos Gomez 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] David Schlangen and Gabriel 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] Ethan O. Selfridge, Peter A Heeman, Iker Arizmendi, and Jason D Williams, "Demonstrating the incremental interaction manager in an end-to-end "lets go!" dialogue system," in *IEEE Workshop on Spoken Language Technology*, 2012.

[6] Nina Dethlefs, Helen Hastie, Verena Rieser, and Oliver 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] Timo Baumann, *Incremental spoken dialogue processing: architecture and lower-level components*, Ph.D. thesis, Bielefeld University, 2013.

[8] Julia Hirschberg, "Communication and prosody: Functional aspects of prosody," *Speech Communication*, vol. 36, no. 1–2, pp. 31–43, 2002.

[9] Janet Hitzeman, Alan W Black, Paul Taylor, Chris Mellish, and Jon Oberlander, "On the use of automatically generated discourse-level information in a concept-to-speech synthesis system," in *Proceedings of the International Conference on Spoken Language Processing (ICSLP98)*, 1998.

[10] Seiya Takada, Yuji Yagi, Keikichi Hirose, and Nobuaki Minematsu, "A framework of reply speech generation for concept-to-speech conversion in spoken dialogue systems.," in *INTERSPEECH*, 2007, pp. 1286–1289.

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

[12] Paul 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.

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

[14] Laurie Hiyakumoto, Scott Prevost, and Justine Cassell, "Semantic and discourse information for text-to-speech intonation," in *ACL Workshop on Concept-to-Speech Technology*, 1997, pp. 47–56.

[15] Shimei Pan, *Prosody modeling in concept-to-speech generation*, Ph.D. thesis, Columbia University, 2002.

[16] Markus Schnell and Rüdiger Hoffmann, "What concept-to-speech can gain for prosody.," in *INTERSPEECH*, 2004.

[17] Heiga Zen, Keiichi Tokuda, and Alan W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[18] Simon King, Alan W Black, Keiichi Tokuda, and Kishore Prahallad, "The Blizzard Challenge 2013," 2013.

[19] Leonardo Badino, J Sebastian Andersson, Junichi Yamagishi, and Robert AJ Clark, "Identification of contrast and its emphatic realization in HMM-based speech synthesis," in *INTERSPEECH*, 2009.

[20] Kai Yu, François Mairesse, and Steve Young, "Word-level emphasis modelling in HMM-based speech synthesis," in *ICASSP*. IEEE, 2010, pp. 4238–4241.

[21] Kumiko Morizane, Keigo Nakamura, Tomoki Toda, Hiroshi Saruwatari, and Kiyohiro Shikano, "Emphasized speech synthesis based on hidden markov models," in *Oriental COCOSDA International Conference on Speech Database and Assessments*. IEEE, 2009, pp. 76–81.

[22] Kai Yu, Heiga Zen, François Mairesse, and Steve 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.

[23] Fanbo Meng, Zhiyong Wu, Jia Jia, Helen Meng, and Lianhong Cai, "Synthesizing english emphatic speech for multimodal corrective feedback in computer-aided pronunciation training," *Multimedia Tools and Applications*, pp. 1–27, 2013.

[24] Yu Maeno, Takashi Nose, Takao Kobayashi, Tomoki Koriyama, Yusuke Ijima, Hideharu Nakajima, Hideyuki Mizuno, and Osamu Yoshioka, "HMM-based expressive speech synthesis based on phrase-level F0 context labeling," in *ICASSP*, 2013, pp. 7859–7863.

[25] Junichi Yamagishi, Takashi Masuko, and Takao 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.

[26] Langzhou Chen, Mark JF Gales, Vincent Wan, Javier Latorre, and Masami Akamine, "Exploring rich expressive information from audiobook data using cluster adaptive training.," in *INTERSPEECH*, 2012.

[27] Oliver Lemon, "The CLASSiC project," 2011.

[28] Pirros Tsiakoulis, Milica Gašić, Matthew Henderson, Joaquin Planells-Lerma, Jorge Prombonas, Blaise Thomson, Kai Yu, Steve Young, and Eli 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] Alan W Black and Kevin A Lenzo, "Building synthetic voices," 2003.

[30] Simon King and Vasilis Karaiskos, "The Blizzard Challenge 2010," 2010.

[31] Kai Yu and Steve 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.